

Delay-Optimal Scheduling for Queueing Systems with Switching Overhead

Ping-Chun Hsieh · I-Hong Hou · Xi Liu

Received: date / Accepted: date

Abstract We study the scheduling policies for asymptotically optimal delay in queueing systems with switching overhead. Such systems consist of a single server that serves multiple queues, and some capacity is lost whenever the server switches to serve a different set of queues. The capacity loss due to this switching overhead can be significant in many emerging applications, and needs to be explicitly addressed in the design of scheduling policies. For example, in 60GHz wireless networks with directional antennas, base stations need to train and reconfigure their beam patterns whenever they switch from one client to another. Considerable switching overhead can also be observed in many other queueing systems such as transportation networks and manufacturing systems. While the celebrated Max-Weight policy achieves asymptotically optimal average delay for systems without switching overhead, it fails to preserve throughput-optimality, let alone delay-optimality, when switching overhead is taken into account. We propose a class of Biased Max-Weight scheduling policies that explicitly takes switching overhead into account. The Biased Max-Weight policy can use either queue length or head-of-line waiting time as an indicator of the system status. We prove that our policies not only are throughput-optimal, but also can be made arbitrarily close to the asymptotic lower bound on average delay. To validate the performance of the proposed policies, we provide extensive simulation with various system topologies and different traffic patterns. We show that the proposed policies indeed achieve much better delay performance than that of the state-of-the-art policy.

Keywords Delay-optimality · Scheduling · Switching overhead · Max-Weight · Throughput-optimality · Stability

Mathematics Subject Classification (2010) 60K25 · 68M20 · 90B22

Ping-Chun Hsieh · I-Hong Hou · Xi Liu
Department of ECE, Texas A&M University, College Station, Texas 77843-3128, USA.
E-mail: {pingchun.hsieh, ihou, xiliu}@tamu.edu

1 Introduction

Design of scheduling policies is one of the most critical parts in achieving good performance for queueing systems. Based on the system information, such as queue backlog or instantaneous service rate, the server dynamically switches between different scheduling decisions. The delay required for each switch is traditionally assumed to be minimal compared to the service time of each job, and therefore can be neglected. However, for applications that require dynamic tuning or strict safety guarantees, this switching overhead has to be explicitly addressed with caution. Wireless networking on the 60 GHz band [11, 25, 28], for example, is inherently featured by the switching overhead. Signal propagation on the 60 GHz band suffers from much more serious attenuation than on the widely-used 2.4 GHz or 5 GHz spectrum. To cope with this increased attenuation, beamforming techniques based on directional antennas have been widely applied to mitigate this problem [24]. However, as stated in [21], to reconfigure the beam direction, the beam switching procedure can take up to several hundred microseconds, which is about the transmission time of a packet. While this switching overhead is usually overlooked in most wireless communication networks, this beam-switching latency needs to be explicitly addressed in directional-antenna applications.

Another example is the traffic signal control for signalized intersections. During each signal phase transition, the traffic signal controller has to undergo a yellow-to-red period and an all-red period to guarantee safety. Without proper scheduling, the switching overhead can greatly reduce the intersection capacity [12]. Existing studies have shown that the conventional fixed time scheduling policy can result in a significant amount of capacity loss. [2, 33]. Moreover, this switching overhead is expected to be even larger in mixed transportation networks with both human-driven and autonomous vehicles [17, 19]. Recently, there has been growing interest in designing scheduling algorithms to achieve maximum traffic throughput in transportation networks [33, 35]. Meanwhile, more traffic-responsive scheduling policies are now applicable due to the recent progress in both vehicle-to-infrastructure and vehicle-to-vehicle communication. Indeed, to approach maximum throughput in real traffic scenarios, the influence of switching overhead on network capacity has to be incorporated and overcome in scheduling design.

The effect of switching overhead is also critical in many other applications, such as multi-thread processors [8], passive optical networks [20], and manufacturing systems [27]. As a classic example, polling system is one widely studied queueing model that incorporates switching overhead. In such systems, the server serve the queues in cyclic order, and the server requires finite time to go from one queue to the next queue. The polling system model has been applied in a variety of applications, such as communication networks [1] and manufacturing systems [27]. The survey of the polling model and summary of the theoretical results can be found in [18, 29, 30, 34]. Despite the progress in the study of the polling model, there are still a vast variety of applications that require multiple queues to be served simultaneously while the polling model assumes only one queue is served at each time.

For queueing systems that allow simultaneous service of multiple queues, a class of scheduling policies named the Max-Weight policies, introduced by Tassiulas and

Ephremides in [31], has been shown to achieve optimal throughput for the multi-hop queueing systems. Later on, there are several extended works of the Max-Weight policies, such as [10, 32]. In addition to throughput-optimality, the Max-Weight policy has also been shown to achieve good delay performance in different queueing systems [9, 13, 15, 16, 22]. To obtain delay bound for the Max-Weight scheduling, one common technique is to set the drift of a Lyapunov function to zero, as in [9, 13, 15, 16, 22]. Besides, Eryilmaz and Srikant [9] derives an asymptotically tight upper bound on queue length for the Max-Weight policy. However, none of the above literature considers switching overhead incurred by the transition between different schedules. In fact, Max-Weight policy fails to preserve throughput-optimality when the switching overhead is considered. The reason is that Max-Weight policy may suffer from excessive switching and therefore waste a significant portion of time on switching.

To remedy the instability issue of Max-Weight policy, Armony and Bambos [3] propose cone policies and batch policies and prove that the two families of policies are throughput-optimal with non-zero switching overhead. Later on, Hung and Chang [14] propose an extended version of Dynamic Cone policy to reduce the complexity and improve delay performance of the conventional cone policies. Chan *et al.* [6] proposes the Maximum-Weight-Matching with hysteresis (MWM-H) policy to achieve optimal throughput with deterministic service rates as well as to properly distribute the queue backlog when the queues become overloaded. Besides, Celik *et al.* [4] propose the Variable Frame-Based Max-Weight (VFMW) policy which incorporate frame structure to avoid frequent switching, and show that the policy preserves throughput-optimality with nonzero switching overhead for interference networks. However, all of the above studies focus only on throughput-optimality and do not provide any theoretical bound on delay performance. In this paper, we regard the VFMW policy in [4] as the state-of-the-art policy and the VFMW policy serves as a reference for the comparison of delay performance. For the delay analysis of queueing systems with switching overhead, Perkins and Kumar [26] propose a class of exhaustive policies to achieve asymptotically optimal delay for single-server queueing systems with deterministic arrival and service processes. However, it is not even clear whether the exhaustive-type policies can also achieve optimal throughput in systems where multiple queues can be served simultaneously.

In this paper, we propose a class of scheduling policies called *Biased Max-Weight* (BMW) policies to achieve both throughput-optimality and delay-optimality in the presence of switching overhead. The weight of each schedule can be defined by either the queue backlog or the waiting time of the head-of-line (HOL) job. Like the Max-Weight policy, the BMW policy still makes scheduling decisions based on the weight of each schedule, but it has a bias towards the current schedule. In other words, the server makes a switch only when the new schedule is better than the previous schedule by an enough margin. The BMW policies share similar design philosophy with the VFMW policy and the MWM-H policy to avoid excessive switching according to the current queue status. Different from those prior works, we are able to characterize the queue length statistics and therefore the average delay. To achieve this goal, we first prove that the BMW scheduling policy is throughput-optimal by showing that the underlying Markov chain is positive recurrent. We further prove that the BMW

policy achieves the asymptotically tight queue length bound by choosing parameter used by BMW arbitrarily close to zero. Through extensive simulation, we show that the BMW policy still provides good delay performance when the server may serve multiple queues simultaneously, subject to some conflicting constraints imposed by the system.

In addition to average delay, the per-queue delay is also crucial for many applications, such as transportation networks. We further compare the queue-length-based BMW (Q-BMW) and the waiting-time-based BMW (W-BMW) in terms of per-queue average delay. Under the Q-BMW policy, each queue has about the same average queue length. By Little's law, the queue with a smaller mean arrival rate is expected to experience a larger average delay. On the other hand, the W-BMW policy follows the same switching scheme as the Q-BMW policy, but with the HOL waiting times as the state variables. Since the HOL waiting time serves as a good indicator of per-queue average delay, the W-BMW can avoid holding up the jobs for too long in the queues with lighter incoming traffic. Simulation results show that the W-BMW policy indeed achieves better fairness than the Q-BMW policy without sacrificing system-wide average delay.

The rest of the paper is organized as follows. Section 2 describes the system model as well as the notations in use. Section 3 provides mathematical preliminaries of queue stability and delay-optimality and also describes the fundamental dilemma in achieving delay-optimality. Section 4 describes the throughput-optimality of the Q-BMW policy. Then, Section 5 provides an asymptotic upper bound on average delay under Q-BMW policy. For the W-BMW policy, Section 6 shows that it achieves throughput-optimality and the same delay upper bound as the Q-BMW policy. Section 7 provides extensive simulation results of queueing networks with different constraints. Finally, Section 8 concludes the paper.

2 Notation and System Models

2.1 Network Model

We consider a time-slotted switched queueing system with one centralized server and $N \in \mathbb{N}_0$ (\mathbb{N}_0 is a shorthand for $\mathbb{N} \cup \{0\}$) parallel queues, which are indexed by $\mathcal{N} = \{1, 2, \dots, N\}$. Time slots are indexed by $t \in \mathbb{N}_0$. Each queue is associated with an exogenous traffic stream. Arriving jobs are first buffered in the queue and leave the system right after the service is completed. The server may be able to serve multiple queues simultaneously. A set of queues that can be served simultaneously is called a *feasible schedule*. We represent each feasible schedule by an N -dimensional binary vector $\mathbf{I} = (I_1, I_2, \dots, I_N)$, where I_i is the indicator function of whether queue i is included in the schedule. Throughout the paper, we focus on *maximal* feasible schedules: a feasible schedule is *maximal* if no additional queue can be added to the schedule. With a little abuse of notation, we use $|\mathbf{I}|$ to denote the number of queues included in the schedule \mathbf{I} . Suppose there are J maximal feasible schedules denoted by $\mathcal{I} = \{\mathbf{I}^{(1)}, \dots, \mathbf{I}^{(J)}\}$. In each time slot t , the server selects a maximal feasible schedule $\mathbf{I}(t) \in \mathcal{I}$ according to its scheduling policy η .

When the server switches from one schedule to serving another schedule, it needs to spend $T_s \in \mathbb{N}$ slots on preparing for the transition before working on the new schedule. The delay T_s reflects the switching overhead which is usually overlooked in the queueing systems but needs to be explicitly addressed in many applications, such as directional-antenna systems as well as transportation systems described in Section 1. Therefore, there are two operation modes for the server: ACTIVE mode and SWITCH mode. We let $M(t)$ be the indicator function of whether the server is in ACTIVE mode at time t . We use t_k to denote the time when the server makes a switch for the k -th time, and set $t_0 = 0$. The time between two consecutive switches is called an *interval*. Let $T_k := t_{k+1} - t_k$ denote the length of the k -th interval, for all $k \in \mathbb{N}_0$. Note that T_k reflects how frequently the server is switching between different schedules.

Throughout the paper, we use $(\mathcal{N}, \mathcal{I}, T_s)$ to denote a queueing system described in the above.

2.2 Traffic Model

We model the arrival process $\{A_i(t)\}_t$ of each queue i by a sequence of independent and identically distributed (i.i.d.) non-negative random variables $A_i(t) \in \mathbb{N}_0$ with $\mathbb{E}[A_i(t)] = \lambda_i$ for all $t \geq 0$. We further assume that $A_i(t)$ is upper bounded, i.e. there exists a finite constant $A_{\max} > 0$ such that $A_i(t) \leq A_{\max}$ for every $t \geq 0$. Similarly, the service process $\{S_i(t)\}_t$ of each queue i is modeled by a sequence of i.i.d. non-negative random variable $S_i(t) \in \mathbb{N}_0$ with $\mathbb{E}[S_i(t)] = \mu_i$ and $S_i(t) \leq S_{\max}$, for all $t \geq 0$. We also assume that the server does not collect any information about the instantaneous service rates. The reason is that with non-zero switching overhead the server is not able to exploit the time-varying service when the service rates are independent across time. Moreover, let $\rho_i := \lambda_i / \mu_i$ be the *normalized traffic load* of queue i . For every queue i , each arriving job is labeled with a unique sequence number. For each queue i , the sequence numbers start from 1 and would indicate the arrival order of the jobs. At each time t , we use $\varphi_i(t)$ to denote the sequence number of the latest completed job of queue i before time t . Let $\{V_i(m)\}_{m \geq 0}$ be the inter-arrival time process of queue i , where $V_i(m)$ denote the inter-arrival time between the two jobs with sequence numbers m and $m + 1$ in queue i .

To simplify notations in later sections, we use boldface letters $\mathbf{A}, \mathbf{S}, \boldsymbol{\lambda}, \boldsymbol{\mu}$ and $\boldsymbol{\rho}$ to denote the N -dimensional vectors of the arrivals, services, mean arrival rates, mean service rates, and normalized traffic loads, respectively. We also use $\lambda_{\max}, \lambda_{\min}, \mu_{\max}$ and μ_{\min} as the shorthands of the maximum mean arrival rate, minimum mean arrival rate, maximum mean service rate, and minimum mean service rate among all the queues, respectively.

2.3 Queue Dynamics

Let $Q_i(t) \in \mathbb{N}_0$ be the number of jobs buffered in queue i at time t . We assume $Q_i(0) = 0$ for every queue i . Define $\hat{S}_i(t) := \min\{Q_i(t), M(t)I_i(t)S_i(t)\}$ to be the

amount of service actually used by each queue i at time t . Throughout this paper, we consider the store-and-forward queueing model, i.e., for each queue i ,

$$Q_i(t+1) = Q_i(t) - \hat{S}_i(t) + A_i(t), \quad \forall t \geq 0. \quad (1)$$

For simplicity, we use $\mathbf{Q}(t) = (Q_1(t), \dots, Q_N(t)) \in \mathbb{N}_0^N$ to denote the queue length status of the system at time t . Moreover, in our model the queue length process $\{\mathbf{Q}(t)\}_{t \geq 0}$ form a discrete-time Markov chain on a countable state space \mathbb{N}_0^N . The total queue length of the system can be written as $\mathbf{1}^T \mathbf{Q}(t)$, where $\mathbf{1}^T$ denotes the N -dimensional all-ones row vector.

Let $W_i(t) \in \mathbb{N}_0$ be the waiting time of the head-of-line (HOL) job of queue i and $\mathbf{W}(t) = (W_1(t), \dots, W_N(t)) \in \mathbb{N}_0^N$ be the corresponding HOL waiting time vector. Then, the HOL waiting time can be updated as follows:

$$W_i(t+1) = \max \left\{ 0, \left(W_i(t) - \sum_{j=1}^{\hat{S}_i(t)} V_i(\varphi(t) + j) \right) \right\}. \quad (2)$$

Similar to the queue length process, the HOL waiting time process $\{\mathbf{W}(t)\}_{t \geq 0}$ also form a discrete-time Markov chain on a countable state space \mathbb{N}_0^N .

3 Preliminaries

3.1 Capacity Region and a Lower-Bound for Delay

In preparation for the discussion of delay performance, we first introduce the fundamental concepts of queue stability, throughput-optimality, and delay-optimality. First, we formally introduce one commonly used definition of queue stability.

Definition 1 (*Strong stability*) The queueing system is *strongly stable* under a scheduling policy if

$$\limsup_{t \rightarrow \infty} \frac{\sum_{\tau=0}^{t-1} \mathbb{E}[\mathbf{1}^T \mathbf{Q}(\tau)]}{t} < \infty. \quad \square \quad (3)$$

Based on the above definitions, we can classify the arrival rate vectors by queue stability and define the capacity region.

Definition 2 (*Admissible arrival rates*) An arrival rate vector $\lambda = (\lambda_1, \dots, \lambda_N)$ is said to be *admissible* if there exists a scheduling policy under which the queueing system is strongly stable. \square

Definition 3 (*Capacity region*) The *capacity region* $\Lambda \subset \mathbb{R}_+^N$ of the system is defined as the closure of the set that consists of all the admissible arrival rate vectors. \square

The following lemma shows that the capacity region can be fully characterized by the normalized traffic load vector and the maximal feasible schedules.

Lemma 1 (*Characterization of capacity region*) For any queueing system described in Section 2, given the mean service rate vector μ , the capacity region can be characterized as

$$\Lambda = \left\{ \lambda \mid \exists \beta \geq 0 \text{ with } \sum_{j=1}^J \beta_j \leq 1 \text{ such that } \rho \leq \sum_{j=1}^J \beta_j \mathbf{I}^{(j)} \right\}. \quad (4)$$

Proof This is a direct result of Theorem 1 in [5]. \square

Here, we introduce the notion of throughput-optimality:

Definition 4 (*Throughput-optimality*) A scheduling policy η is said to be *throughput-optimal* if for any interior point λ of Λ , the system is strongly stable under η . \square

Given λ , μ , and \mathcal{I} , we can further describe the traffic load of the whole system.

Definition 5 (*Utilization factor*) Given the queueing system described in Section 2, we define the *utilization factor* as

$$\beta^* := \min_{\beta: \sum_{j=1}^J \beta_j \mathbf{I}^{(j)} \geq \rho} \mathbf{1}^T \beta. \quad (5)$$

For convenience, we also define $\epsilon^* := 1 - \beta^*$, which reflects the "distance" from the boundary of the capacity region. \square

From the study in [9], the average delay of a queueing system is closely related to ϵ^* . A useful lower bound for the average delay is provided here as an easy reference.

Lemma 2 (*Lower bound on queue length without switching overhead*) Given a stable queueing system $\mathcal{Q} = (\mathcal{N}, \mathcal{I}, T_s)$ described in Section 2 with queue length process $\{\mathbf{Q}(t)\}_t$ and $T_s = 0$, under any scheduling policy η , the expected total queue length in steady state scales as

$$\mathbb{E}[\mathbf{1}^T \mathbf{Q}(t)] = \Omega(1/\epsilon^*). \quad (6)$$

Proof This is a direct result of Lemma 6 in [9]. \square

Remark 1 By Little's law, (6) implies that the total average delay also scales as $\Omega(1/\epsilon^*)$ for systems without switching overhead.

Remark 2 In [9], the lower bound (6) is obtained by constructing a hypothetical single-queue system from the original multi-queue system, whose total queue length is larger in stochastic ordering than that of the constructed single-queue system. Following this argument, we can also derive a similar lower bound on expected queue length for the queueing systems with switching overhead.

Corollary 1 (*Lower bound on queue length with switching overhead*) Given a stable queueing system $\mathcal{Q} = (\mathcal{N}, \mathcal{I}, T_s)$ described in Section 2 with queue length process $\{\mathbf{Q}(t)\}_t$ and $T_s > 0$, under any scheduling policy η , the expected total queue length in steady state is

$$\mathbb{E}[\mathbf{1}^T \mathbf{Q}(t)] = \Omega(1/\epsilon^*). \quad (7)$$

Proof Given a queueing system \mathcal{Q} , by following the same procedure as in Lemma 6 of [9], we can construct a single-queue system \mathcal{Q}' with switching overhead T_s . Then we know the queue length process of \mathcal{Q} is larger in stochastic ordering than the queue length process of \mathcal{Q}' . Next, we can construct another single-queue system \mathcal{Q}'' from \mathcal{Q}' but with $T_s = 0$. Then we know the queue length process of \mathcal{Q}' is larger in stochastic ordering than the queue length process of \mathcal{Q}'' . By Lemma 2, we have $\mathbb{E}[\mathbf{1}^T \mathbf{Q}(t)] = \Omega(1/\epsilon^*)$. \square

Based on Corollary 1, we can define asymptotic delay-optimality as follows.

Definition 6 (*Delay-optimality*) A scheduling policy is *delay-optimal* if in steady state, the total average queue length satisfies that

$$\mathbb{E}[\mathbf{1}^T \mathbf{Q}(t)] = O(1/\epsilon^*). \quad (8)$$

In other words, $O(1/\epsilon^*)$ is an asymptotically tight delay upper bound. \square

3.2 The Variable-Frame Max-Weight Scheduling Policy

As discussed in Section 1, despite the progress in throughput-optimal scheduling for systems with switching overhead, it is still not clear how to achieve optimal delay performance for such queueing systems with stochastic arrival and service processes. In the prior work [4], the Variable-Frame Max-Weight (VFMW) policy has been proposed to achieve throughput optimality queueing systems with switching overhead. Under the VFMW policy, we need to determine a sublinear function for calculating frame size such that the server stays with the same schedule till the end of the frame. While the frame size function has no effect on throughput-optimality (as long as it is sublinear), it can indeed greatly affect the delay performance. In the example discussed in [4], the frame function is chosen to be $(\sum_i Q_i(t))^\alpha$, where α is between 0 and 1. Besides, [4] also suggests that α should be chosen as close to 1 as possible based on their simulation results. However, the α value with the smallest delay can actually differ in different scenarios. This phenomenon can be easily observed through simulation as follows.

For example, we consider a single-server system of 4 queues with Bernoulli arrival and service processes. The switching overhead $T_s = 1$. We provide simulation results of two scenarios with different mean arrival rates and mean service rates in Figure 1(a) and 1(b).

- Scenario I: $\lambda = (0.119, 0.119, 0.119, 0.119)$, $\mu = (0.5, 0.5, 0.5, 0.5)$
- Scenario II: $\lambda = (0.08, 0.25, 0.09, 0.01)$, $\mu = (0.8, 0.5, 0.3, 0.2)$

Note that the utilization factor of both scenarios is 0.95. In Scenario I, the smallest total delay is achieved when α is close to 1. However, the optimal α is about 0.6 in Scenario II. This example demonstrates that there does not exist a fixed value of α that can achieve the optimal delay performance for the VFMW policy. We study the trace files of our simulations and find that the VFMW policy can suffer from poor delay performance from two conflicting factors:

- If α is large, say close to 1, then the frame size grows fast. The VFMW policy may stick to an inefficient feasible schedule for too long and thereby suffers from large delay.
- If α is small, say close to 0, then the frame size is very small for most of the time and becomes very insensitive to the change in queue status. Consequently, the VFMW policy may switch too frequently and gets severely impacted by the switching overhead.

The above arguments highlight the fundamental difficulty in achieving delay-optimality for queueing systems with switching overhead: If a policy switches too frequently, it suffers from too much capacity loss due to switching overhead. On the other hand, if a policy does not switch often enough, it may stay with a schedule that is no longer efficient for too long. In the next section, we propose our online scheduling policy that solves such dilemma.

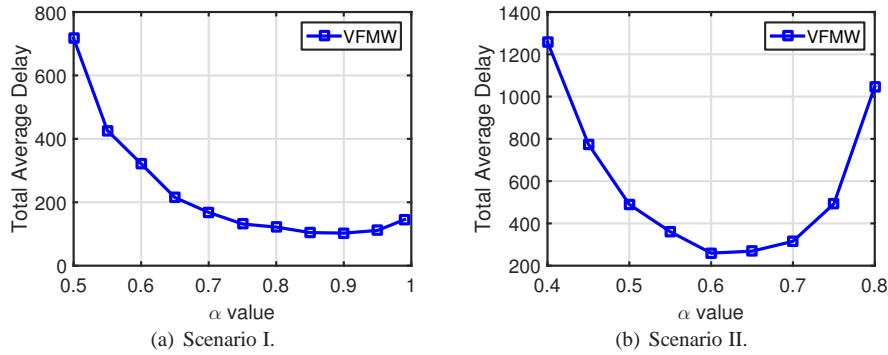


Fig. 1 Average total delay versus different α value under VFMW policy with frame function $(\sum_i Q_i(t))^\alpha$.

4 Throughput-Optimal Scheduling for Systems With Switching Overhead

4.1 Scheduling Policy and Intuitive Description

We propose two types of *Biased Max-Weight scheduling policy* to achieve both throughput-optimality and the asymptotically tight delay bound. We first introduce the queue-length-based Biased Max-Weight (Q-BMW) policy as follows.

Q-BMW policy: Let $F(\cdot) : \mathbb{R}_+^N \rightarrow [1, \infty)$ be a function chosen by the server. At each time t in the k -th interval, if the system satisfies that

$$\left(1 + \frac{T_s}{F(\mathbf{Q}(t_k))}\right) \left(\mathbf{I}(t_k)^T \mathbf{Q}(t)\right) \leq \left(\max_{j: 1 \leq j \leq J} (\mathbf{I}^{(j)})^T \mathbf{Q}(t)\right), \quad (9)$$

then the server makes a switch to serve the schedule with the largest sum of queue lengths at time t (ties are broken arbitrarily). Otherwise, the server stays with the current schedule. \square

Note that the Q-BMW policy does not rely on the frame structure adopted by the VFMW policy. Instead, the Q-BMW policy overcomes the switching overhead by giving an intentional bias to the current schedule. Moreover, the Q-BMW keeps checking the condition of (9) in each time slot. Intuitively, the Q-BMW policy avoids the dilemma highlighted in the previous section because:

- Since the Q-BMW policy checks (9) in each time slot, it cannot stick to an inefficient schedule for too long, regardless of the choice of the function $F(\cdot)$.
- Since the Q-BMW favors the current schedule, it can avoid switching too frequently as long as the bias to the current schedule is not too small. This suggests that one should choose a function $F(\cdot)$ that increases very slowly with $\mathbf{Q}(t)$.

4.2 Throughput-Optimality of the Q-BMW Scheduling

To show that the Q-BMW is throughput-optimal, we first introduce a lower bound on T_k in the following lemma.

Lemma 3 *Suppose the server can serve at most K queues at a time. Under the Q-BMW policy, for any $k \geq 0$ and for every sample path, in the k -th interval we have*

$$T_k \geq C_0(\mathbf{1}^T \mathbf{Q}(t_k))/F(\mathbf{Q}(t_k)), \quad (10)$$

where $C_0 = T_s/(NK(A_{\max} + (1 + T_s)S_{\max}))$. \square

Proof Suppose at time $t = t_k + \tau$, $\tau > 0$ in the k -th interval, the server enters SWITCH mode and starts switching. Then, there exists some schedule $\mathbf{I}^{(m)} \neq \mathbf{I}(t_k)$ such that

$$\left(1 + \frac{T_s}{F(\mathbf{Q}(t_k))}\right) \left(\mathbf{I}(t_k)^T \mathbf{Q}(t_k + \tau)\right) \leq \left(\mathbf{I}^{(m)}^T \mathbf{Q}(t_k + \tau)\right). \quad (11)$$

Moreover, by the boundedness of the arrival processes, we know

$$(\mathbf{I}^{(m)})^T (\mathbf{Q}(t_k) + \tau A_{\max} \mathbf{1}) \geq (\mathbf{I}^{(m)})^T \mathbf{Q}(t_k + \tau). \quad (12)$$

From (11) and (12), we have

$$(\mathbf{I}^{(m)})^T (\mathbf{Q}(t_k) + \tau A_{\max} \mathbf{1}) \geq (\mathbf{I}(t_k)^T \mathbf{Q}(t_k + \tau)) \left(1 + \frac{T_s}{F(\mathbf{Q}(t_k))}\right) \quad (13)$$

$$\geq (\mathbf{I}(t_k)^T (\mathbf{Q}(t_k) - \tau S_{\max} \mathbf{1})) \left(1 + \frac{T_s}{F(\mathbf{Q}(t_k))}\right) \quad (14)$$

Next, we rearrange the above equations as

$$K(A_{\max} + (1 + T_s)S_{\max})\tau \geq \mathbf{I}(t_k)^T \mathbf{Q}(t_k) - (\mathbf{I}^{(m)})^T \mathbf{Q}(t_k) + \frac{T_s \mathbf{I}(t_k)^T \mathbf{Q}(t_k)}{F(\mathbf{Q}(t_k))} \quad (15)$$

$$\geq \frac{T_s \mathbf{I}(t_k)^T \mathbf{Q}(t_k)}{F(\mathbf{Q}(t_k))} \quad (16)$$

$$\geq \frac{T_s \mathbf{1}^T \mathbf{Q}(t_k)}{N \cdot F(\mathbf{Q}(t_k))} \quad (17)$$

Hence, we can get the lower bound on T_k as

$$T_k \geq \frac{T_s \cdot \mathbf{1}^T \mathbf{Q}(t_k)}{NK(A_{\max} + (1 + T_s)S_{\max}) \cdot F(\mathbf{Q}(t_k))}. \quad (18)$$

□

Lemma 3 provides important insight in choosing a proper $F(\mathbf{Q}(t_k))$. We now state the main theorem of throughput-optimality as follows.

Theorem 1 *If we choose $F(\mathbf{Q}(t_k)) = \max\{1, (\mathbf{1}^T \mathbf{Q}(t_k))^\alpha\}$ with $\alpha \in (0, 1)$, then the Q-BMW policy is throughput-optimal. Moreover, the underlying Markov chain induced by the queue length process $\{\mathbf{Q}(t)\}_{t \geq 0}$ is positive recurrent. □*

Proof The complete proof is provided in Appendix 1. In summary, for any system with mean arrival rates vector λ in the interior of the capacity region Λ , we show that the queueing system is strongly stable by applying the Lyapunov drift framework. That is, we utilize a quadratic Lyapunov function and show that the expected Lyapunov drift is negative. Different from the one-step drift which is often used in the queueing systems without switching overhead, we choose a hypothetical observation window and show that the multi-step Lyapunov drift across the window is negative. Based on the lower bound on T_k given by (18), the effect of switching overhead is amortized over the whole observation window. □

5 Asymptotically Tight Queue Length Bound under Q-BMW Scheduling

In this section, we focus on systems where the server can serve at most one queue at a time, that is, $|\mathbf{I}| = 1$, for all feasible schedule \mathbf{I} . We show that Q-BMW is nearly delay-optimal when $\alpha \downarrow 0$ by proving the following theorem:

Theorem 2 *For any queueing system $\mathcal{Q} = (\mathcal{N}, \mathcal{I}, T_s)$ described in Section 2 where the server can serve at most one queue at a time, the Q-BMW scheduling policy provides the following queue length upper bound: there exists some constant $B < \infty$ such that*

$$\lim_{\epsilon^* \downarrow 0} \epsilon^* \mathbb{E}[(\mathbf{1}^T \mathbf{Q}(t))^{1-\alpha}] \leq B. \quad (19)$$

Hence, $\mathbb{E}[(\mathbf{1}^T \mathbf{Q}(t))^{1-\alpha}]$ scales as $O(1/\epsilon^*)$. By choosing α arbitrarily close to 0, the Q-BMW policy achieves the asymptotically tight queue length bound. □

We first introduce some necessary definitions and lemmas for the proof of Theorem 2.

Definition 7 A scheduling policy η is said to be *work-conserving* if the server never serves an empty queue whenever there is an unfinished job in the system. \square

Definition 8 A scheduling policy η is said to be *ergodic* if the Markov chain resulting from η is positive recurrent. \square

Lemma 4 Let $K_{\mathbb{T}}$ be the number of intervals in $[0, \mathbb{T})$. For any queueing system under any ergodic work-conserving policy η , there exists some constant $B_0 < \infty$ such that

$$\limsup_{\mathbb{T} \rightarrow \infty} \frac{\sum_{k=1}^{K_{\mathbb{T}}} T_k}{K_{\mathbb{T}}} \leq \frac{B_0}{\epsilon^*}, \quad (20)$$

almost surely. \square

Proof Let the counting process $Y_i(t)$ be the number of slots for which queue i is served up to t . Moreover, let $\tau_{i,1}, \tau_{i,2}, \dots, \tau_{i,Y_i(t)}$ be these time slots. Let $Z_i(t)$ denote the cumulative time for which the scheduled service for queue i is not fully utilized up to time t , that is,

$$Z_i(t) := \sum_{\tau=0}^{t-1} \mathbb{I}_{\{\hat{S}_i(\tau) > Q_i(\tau)\}}, \quad (21)$$

where $\mathbb{I}_{\{\cdot\}}$ denotes the indicator function of an event. At each time τ , the event that $\{\hat{S}_i(\tau) > Q_i(\tau)\}$ happens only when the queue becomes empty at the beginning of time $\tau + 1$. Since the policy η is work-conserving, at the beginning of slot $\tau + 1$ the server will switch to a new schedule. Therefore, the total cumulative time for which the scheduled service is not fully utilized is at most the same as the number of intervals, that is, for any $\mathbb{T} > 0$,

$$\sum_{i=1}^N Z_i(\mathbb{T}) \leq K_{\mathbb{T}}, \quad (22)$$

where $K_{\mathbb{T}}$ is the number of intervals in $[0, \mathbb{T})$. Since the policy η is work-conserving, we also have

$$\sum_{i=1}^N Y_i(\mathbb{T}) \geq \mathbb{T} - K_{\mathbb{T}} T_s. \quad (23)$$

Moreover, we have for each queue i ,

$$S_{\max} Z_i(\mathbb{T}) \geq \sum_{m=1}^{Y_i(\mathbb{T})} S_i(\tau_{i,m}) - \sum_{\tau=0}^{\mathbb{T}-1} A_i(\tau). \quad (24)$$

The right-hand side of (24) represents the cumulative service that is not fully utilized by queue i up to \mathbb{T} . After dividing both sides of (24) by \mathbb{T} , we have

$$\frac{S_{\max} Z_i(\mathbb{T})}{\mathbb{T}} \geq \frac{\sum_{m=1}^{Y_i(\mathbb{T})} S_i(\tau_{i,m})}{\mathbb{T}} - \frac{\sum_{\tau=0}^{\mathbb{T}-1} A_i(\tau)}{\mathbb{T}} \quad (25)$$

$$= \frac{Y_i(\mathbb{T})}{\mathbb{T}} \frac{\sum_{m=1}^{Y_i(\mathbb{T})} S_i(\tau_{i,m})}{Y_i(\mathbb{T})} - \frac{\sum_{\tau=0}^{\mathbb{T}-1} A_i(\tau)}{\mathbb{T}} \quad (26)$$

Since the policy η is ergodic and thereby the Markov chain resulting from η is positive recurrent, then $\lim_{\mathbb{T} \rightarrow \infty} \frac{Y_i(\mathbb{T})}{\mathbb{T}}$ exists, for every queue i . By letting $\mathbb{T} \rightarrow \infty$, we have

$$\liminf_{\mathbb{T} \rightarrow \infty} \frac{S_{\max} Z_i(\mathbb{T})}{\mathbb{T}} \geq \liminf_{\mathbb{T} \rightarrow \infty} \frac{Y_i(\mathbb{T})}{\mathbb{T}} \frac{\sum_{m=1}^{Y_i(\mathbb{T})} S_i(\tau_{i,m})}{Y_i(\mathbb{T})} - \limsup_{\mathbb{T} \rightarrow \infty} \frac{\sum_{\tau=0}^{\mathbb{T}-1} A_i(\tau)}{\mathbb{T}}. \quad (27)$$

Note that $\lim_{\mathbb{T} \rightarrow \infty} Y_i(\mathbb{T}) \rightarrow \infty$ since the queue i cannot be stable if otherwise. By the Strong Law of Large Numbers, we have $\lim_{\mathbb{T} \rightarrow \infty} \frac{\sum_{\tau=0}^{\mathbb{T}-1} A_i(\tau)}{\mathbb{T}} = \lambda_i$ and $\lim_{\mathbb{T} \rightarrow \infty} \frac{\sum_{m=1}^{Y_i(\mathbb{T})} S_i(\tau_{i,m})}{Y_i(\mathbb{T})} = \mu_i$, for every queue i . Therefore, (27) can be written as

$$\frac{S_{\max}}{\mu_i} \cdot \liminf_{\mathbb{T} \rightarrow \infty} \frac{Z_i(\mathbb{T})}{\mathbb{T}} \geq \lim_{\mathbb{T} \rightarrow \infty} \frac{Y_i(\mathbb{T})}{\mathbb{T}} - \rho_i. \quad (28)$$

By summing (28) over all i , we have

$$\frac{NS_{\max}}{\mu_{\min}} \cdot \liminf_{\mathbb{T} \rightarrow \infty} \frac{\sum_{i=1}^N Z_i(\mathbb{T})}{\mathbb{T}} \geq \lim_{\mathbb{T} \rightarrow \infty} \frac{\sum_{i=1}^N Y_i(\mathbb{T})}{\mathbb{T}} - \sum_{i=1}^N \rho_i \quad (29)$$

$$\geq 1 - \liminf_{\mathbb{T} \rightarrow \infty} \frac{K_{\mathbb{T}} T_s}{\mathbb{T}} - \sum_{i=1}^N \rho_i \quad (30)$$

$$= \epsilon^* - \liminf_{\mathbb{T} \rightarrow \infty} \frac{K_{\mathbb{T}} T_s}{\mathbb{T}}, \quad (31)$$

where (30) holds from the inequality (23). By using (22), we then have

$$\frac{NS_{\max}}{\mu_{\min}} \cdot \liminf_{\mathbb{T} \rightarrow \infty} \frac{K_{\mathbb{T}}}{\mathbb{T}} \geq \epsilon^* - \liminf_{\mathbb{T} \rightarrow \infty} \frac{K_{\mathbb{T}} T_s}{\mathbb{T}}. \quad (32)$$

Therefore, we obtain that $\liminf_{\mathbb{T} \rightarrow \infty} \frac{K_{\mathbb{T}}}{\mathbb{T}} \geq \epsilon^* (T_s + \frac{NS_{\max}}{\mu_{\min}})^{-1}$, almost surely. Equivalently, we have

$$\limsup_{\mathbb{T} \rightarrow \infty} \frac{\sum_{k=1}^{K_{\mathbb{T}}} T_k}{K_{\mathbb{T}}} \leq \frac{T_s + \frac{NS_{\max}}{\mu_{\min}}}{\epsilon^*}, \quad (33)$$

almost surely. \square

Lemma 5 For any queueing system $\mathcal{Q} = (\mathcal{N}, \mathcal{I}, T_s)$ described in Section 2 where the server can serve at most one queue at a time, the Q -BMW scheduling policy is work-conserving. \square

Proof Under the Q-BMW policy, if the scheduled queue becomes empty, that is, $\mathbf{I}(t)^T \mathbf{Q}(t) = 0$, and if there still exists another non-empty queue, then the switching condition (9) should be triggered. Therefore, the Q-BMW policy never idles when there is still a job in the system. \square

Theorem 3 For any queueing system $\mathcal{Q} = (\mathcal{N}, \mathcal{I}, T_s)$ described in Section 2 where the server can serve at most one queue at a time, under the Q-BMW scheduling policy, there exists some constant $B_0 < \infty$ such that

$$\limsup_{T \rightarrow \infty} \frac{\sum_{k=1}^{K_T} T_k}{K_T} \leq \frac{B_0}{\epsilon^*}, \quad (34)$$

almost surely. Moreover, if the system is strongly stable and therefore the underlying Markov chain is positive recurrent, then we also have

$$\lim_{\epsilon^* \downarrow 0} \epsilon^* \mathbb{E}[T_k] \leq B_0. \quad (35)$$

\square

Proof This is a direct result of Lemma 4 and Lemma 5. \square

The following lemma shows that to derive the queue length bound in steady state, we can consider only the queue length at the beginning of each interval.

Lemma 6 Given $\gamma \in (0, 1]$, in steady state, if there exists some positive constant $B_0 < \infty$ such that at the beginning of any interval

$$\lim_{\epsilon^* \downarrow 0} \epsilon^* \mathbb{E}[(\mathbf{1}^T \mathbf{Q}(t_k))^\gamma] \leq B_0, \quad (36)$$

then there also exists a positive constant $B_1 < \infty$ such that in any time slot t

$$\lim_{\epsilon^* \downarrow 0} \epsilon^* \mathbb{E}[(\mathbf{1}^T \mathbf{Q}(t))^\gamma] \leq B_1. \quad (37)$$

\square

Proof For any time slot t in the k -th interval, we have

$$\mathbb{E}[(\mathbf{1}^T \mathbf{Q}(t))^\gamma] \leq \mathbb{E}\left[(\mathbf{1}^T \mathbf{Q}(t_k) + (t - t_k) \sum_{i=1}^N A_{\max})^\gamma\right] \quad (38)$$

$$\leq \mathbb{E}\left[(\mathbf{1}^T \mathbf{Q}(t_k))^\gamma + (t - t_k) \sum_{i=1}^N A_{\max}\right] \quad (39)$$

$$\leq \mathbb{E}\left[(\mathbf{1}^T \mathbf{Q}(t_k))^\gamma + N A_{\max} T_k\right]. \quad (40)$$

Therefore, by Theorem 3, we know that

$$\lim_{\epsilon^* \downarrow 0} \epsilon^* \mathbb{E}[(\mathbf{1}^T \mathbf{Q}(t))^\gamma] \leq \lim_{\epsilon^* \downarrow 0} \epsilon^* \mathbb{E}[(\mathbf{1}^T \mathbf{Q}(t_k))^\gamma] + \lim_{\epsilon^* \downarrow 0} \epsilon^* \mathbb{E}[N A_{\max} T_k] < \infty. \quad (41)$$

This completes the proof. \square

We are now ready to prove Theorem 2.

Proof (Theorem 2) Given $\mathbf{Q}(t_k)$, for any $\tau \geq T_s$, define $\Delta\tilde{\mathbf{Q}}(t_k + \tau) := \mathbf{Q}(t_k + \tau) - \left(\mathbf{Q}(t_k) + \tau\boldsymbol{\lambda} - (\tau - T_s)(\boldsymbol{\mu} \circ \mathbf{I}(t_k))\right)$, where $\boldsymbol{\mu} \circ \mathbf{I}(t_k)$ denotes the element-wise product of $\boldsymbol{\mu}$ and $\mathbf{I}(t_k)$. Note that $\Delta\tilde{\mathbf{Q}}(t_k + \tau)$ represents the "deviation" in queue backlog with stochastic arrival and service processes from that with deterministic arrival rates and service rates. Therefore, at time t_{k+1} , under the Q-BMW policy,

$$\left(1 + \frac{T_s}{F(\mathbf{Q}(t_k))}\right) \left(\mathbf{I}(t_k)^T \left(\mathbf{Q}(t_k) + T_k\boldsymbol{\lambda} - (T_k - T_s)\boldsymbol{\mu} + \Delta\tilde{\mathbf{Q}}(t_{k+1})\right)\right) \quad (42)$$

$$\leq \left(\mathbf{I}(t_{k+1})^T \left(\mathbf{Q}(t_k) + T_k\boldsymbol{\lambda} + \Delta\tilde{\mathbf{Q}}(t_{k+1})\right)\right). \quad (43)$$

Since $\mathbf{I}(t_k)^T \mathbf{Q}(t_k) \geq \mathbf{I}(t_{k+1})^T \mathbf{Q}(t_k)$, (42) and (43) can be rearranged as

$$\frac{T_s \mathbf{I}(t_k)^T \mathbf{Q}(t_k)}{F(\mathbf{Q}(t_k))} \leq \left(1 + \frac{T_s}{F(\mathbf{Q}(t_k))}\right) \left(\left(T_k \mathbf{I}(t_k)^T (\boldsymbol{\mu} - \boldsymbol{\lambda})\right) - \mathbf{I}(t_k)^T \Delta\tilde{\mathbf{Q}}(t_{k+1}) \right) \quad (44)$$

$$+ T_k \mathbf{I}(t_{k+1})^T \boldsymbol{\lambda} + \mathbf{I}(t_{k+1})^T \Delta\tilde{\mathbf{Q}}(t_{k+1}) \quad (45)$$

$$\leq T_k \left(\mu_{\max}(T_s + 1) + \lambda_{\max}\right) + (T_s + 2) \sum_{i=1}^N |\Delta\tilde{Q}_i(t_{k+1})|. \quad (46)$$

By the Functional Law of Iterated Logarithm [7], with probability one we have

$$\Delta\tilde{Q}_i(t_{k+1}) = O(\sqrt{T_k \log \log T_k}), \quad \forall i = 1, \dots, N. \quad (47)$$

Therefore, by choosing $F(\mathbf{Q}(t_k))$ as in Theorem 1, we have

$$\left(\sum_{k=1}^{K_{\mathbb{T}}} \left(\mathbf{1}^T \mathbf{Q}(t_k)\right)^{1-\alpha}\right) \leq N \sum_{k=1}^{K_{\mathbb{T}}} \left((\mu_{\max}(T_s + 1) + \lambda_{\max}) \frac{T_k}{T_s} + O(\sqrt{T_k \log \log T_k}) \right). \quad (48)$$

By dividing both sides of (48) by $K_{\mathbb{T}}$ and using Theorem 3, we know there exists some constant $B_0 < \infty$ such that

$$\lim_{\mathbb{T} \rightarrow \infty} \frac{\sum_{k=1}^{K_{\mathbb{T}}} \left(\mathbf{1}^T \mathbf{Q}(t_k)\right)^{1-\alpha}}{K_{\mathbb{T}}} \leq \frac{B_0}{\epsilon^*}, \quad (49)$$

almost surely. For any $\alpha \in (0, 1)$, by Theorem 1, we know that the Markov chain induced by $\{Q(t)\}_{t \geq 0}$ is positive recurrent and therefore

$$\mathbb{E} \left[\left(\mathbf{1}^T \mathbf{Q}(t_k)\right)^{1-\alpha} \right] = \lim_{\mathbb{T} \rightarrow \infty} \frac{\sum_{k=1}^{K_{\mathbb{T}}} \left(\mathbf{1}^T \mathbf{Q}(t_k)\right)^{1-\alpha}}{K_{\mathbb{T}}}, \quad (50)$$

almost surely. Hence, by Lemma 6 along with (49) and (50), there exists a positive constant $B < \infty$ such that

$$\lim_{\epsilon^* \downarrow 0} \epsilon^* \mathbb{E} \left[\left(\mathbf{1}^T \mathbf{Q}(t)\right)^{1-\alpha} \right] \leq B. \quad (51)$$

By choosing α arbitrarily close to 0, the Q-BMW policy indeed achieves the asymptotically tight queue length bound. \square

6 Waiting-Time-Based Biased Max-Weight Scheduling

6.1 Throughput-Optimality

We extend the framework introduced in Section 4 and Section 5 to the waiting-time-based Biased Max-Weight (W-BMW) scheduling policy. Throughout this section, we relax the assumption that the arrival processes are i.i.d. for all the queues. Instead, we make a mild assumption on the arrival processes: for each queue i , the inter-arrival times $\{V_i(m)\}_{m \geq 0}$ form an i.i.d. sequence and are upper bounded by a constant $V_{\max} < \infty$, almost surely. Note that with this assumption, the following analysis of W-BMW policy also applies to queueing systems with periodic arrivals.

W-BMW policy: Let $G(\cdot) : \mathbb{R}_+^N \rightarrow [1, \infty)$ be a function chosen by the server. At each time t in the k -th interval, if the system satisfies

$$\left(1 + \frac{T_s}{G(\mathbf{W}(t_k))}\right) \left(\mathbf{I}(t_k)^T \mathbf{W}(t)\right) \leq \left(\max_{j: 1 \leq j \leq J} (\mathbf{I}^{(j)})^T \mathbf{W}(t)\right), \quad (52)$$

then the server enters SWITCH mode to prepare for serving the schedule with the largest sum of HOL waiting time at time t (ties are broken arbitrarily). Otherwise, the server stays with the current schedule. \square

Remark 3 Both the Q-BMW and W-BMW have the same performance in terms of throughput-optimality and delay-optimality defined in Section 3. The advantage of W-BMW is that it achieves better fairness than the Q-BMW policy in terms of per-queue average delay, especially when there is a large difference in the arrival rates and service rates between different queues. We will further describe this feature of W-BMW through simulation in Section 7.

Lemma 7 Suppose the server can serve at most K queues at a time. Under the W-BMW policy, for every $k \geq 0$ and for every sample path, in the k -th interval we have

$$T_k \geq C_1 (\mathbf{1}^T \mathbf{W}(t_k)) / F(\mathbf{W}(t_k)), \quad (53)$$

where $C_1 = T_s / (NK(1 + (1 + T_s)S_{\max}V_{\max}))$. \square

Proof Suppose at time $t = t_k + \tau$, $\tau > 0$, the server enters SWITCH mode and starts switching. Then, there exists some schedule $\mathbf{I}^{(m)} \neq \mathbf{I}(t_k)$ such that

$$\left(1 + \frac{T_s}{G(\mathbf{W}(t_k))}\right) \left(\mathbf{I}(t_k)^T \mathbf{W}(t_k + \tau)\right) \leq \left((\mathbf{I}^{(m)})^T \mathbf{W}(t_k + \tau)\right). \quad (54)$$

Moreover, we know

$$(\mathbf{I}^{(m)})^T (\mathbf{W}(t_k) + \tau \mathbf{1}) \geq (\mathbf{I}^{(m)})^T \mathbf{W}(t_k + \tau). \quad (55)$$

From (54) and (55), we have

$$(\mathbf{I}^{(m)})^T (\mathbf{W}(t_k) + \tau \mathbf{1}) \geq (\mathbf{I}(t_k)^T \mathbf{W}(t_k + \tau)) \left(1 + \frac{T_s}{G(\mathbf{W}(t_k))}\right) \quad (56)$$

$$\geq (\mathbf{I}(t_k)^T (\mathbf{W}(t_k) - \tau S_{\max} V_{\max} \mathbf{1})) \left(1 + \frac{T_s}{G(\mathbf{W}(t_k))}\right) \quad (57)$$

Next, we rearrange the above equations as

$$K \left(1 + (1 + T_s) S_{\max} V_{\max}\right) \tau \geq \mathbf{I}(t_k)^T \mathbf{W}(t_k) - (\mathbf{I}^{(m)})^T \mathbf{W}(t_k) + \frac{T_s \mathbf{I}(t_k)^T \mathbf{W}(t_k)}{G(\mathbf{W}(t_k))} \quad (58)$$

$$\geq \frac{T_s \mathbf{I}(t_k)^T \mathbf{W}(t_k)}{G(\mathbf{W}(t_k))} \quad (59)$$

$$\geq \frac{T_s \mathbf{1}^T \mathbf{W}(t_k)}{N \cdot G(\mathbf{W}(t_k))} \quad (60)$$

Hence, we can get the lower bound on T_k :

$$T_k \geq \frac{T_s \cdot \mathbf{1}^T \mathbf{W}(t_k)}{NK(1 + (1 + T_s) S_{\max} V_{\max}) \cdot G(\mathbf{W}(t_k))}. \quad (61)$$

□

Next, we show that W-BMW policy is also throughput-optimal in the following theorem.

Theorem 4 *If we choose $G(\mathbf{W}(t_k)) = \max\{1, (\mathbf{1}^T \mathbf{W}(t_k))^\alpha\}$ with $\alpha \in (0, 1)$, then the W-BMW policy is throughput-optimal. Moreover, the underlying Markov chain induced by the waiting time process $\{\mathbf{W}(t)\}_{t \geq 0}$ is positive recurrent. □*

Proof The proof is provided in Appendix 2. We use the similar technique as in the proof of Theorem 1 to prove that W-BMW is throughput-optimal. □

6.2 Asymptotically Tight Queue Length Bound under the W-BMW Scheduling

As in Section 5, we focus on systems where the server can serve at most one queue at a time. We show that W-BMW is also nearly delay-optimal when $\alpha \downarrow 0$ by proving the following theorem:

Theorem 5 *Suppose the server can serve at most one queue at a time. For any such queueing system $\mathcal{Q} = (\mathcal{N}, \mathcal{I}, T_s)$ and stochastic arrival and service processes as described in Section 2 and Section 6, W-BMW policy provides the following upper bound on queue length: there exists some constant $B < \infty$ such that*

$$\lim_{\epsilon^* \downarrow 0} \epsilon^* \mathbb{E} \left[(\mathbf{1}^T \mathbf{Q}(t))^{1-\alpha} \right] \leq B. \quad (62)$$

Hence, $\mathbb{E}[(1^T \mathbf{Q}(t))^{1-\alpha}]$ scales as $O(1/\epsilon^*)$. By choosing α to be arbitrarily close to 0, the W-BMW policy achieves asymptotically tight queue length bound and hence it is delay-optimal. \square

We introduce some necessary lemmas for the proof of Theorem 5.

Lemma 8 Suppose the server can serve at most one queue at a time. For any queueing system $\mathcal{Q} = (\mathcal{N}, \mathcal{I}, T_s)$ described in Section 2, the W-BMW policy is work-conserving. \square

Proof By definition, $Q_i(t) = 0$ implies $W_i(t) = 0$ for any queue i and any time t . Under the W-BMW policy, if the scheduled queue becomes empty, then we have $\mathbf{I}(t)^T \mathbf{W}(t) = 0$. Meanwhile, if there also exists another non-empty queue, then the switching condition (52) should be triggered. Therefore, the W-BMW policy never idles when there is still a job in the system. \square

Theorem 6 Let $K_{\mathbb{T}}$ be the number of intervals in $[0, \mathbb{T})$. For any queueing system $\mathcal{Q} = (\mathcal{N}, \mathcal{I}, T_s)$ described in Section 2 where the server can serve at most one queue at a time, under the W-BMW policy, there exists some constant $B_0 < \infty$ such that

$$\lim_{\mathbb{T} \rightarrow \infty} \frac{\sum_{k=1}^{K_{\mathbb{T}}} T_k}{K_{\mathbb{T}}} \leq \frac{B_0}{\epsilon^*}, \quad (63)$$

almost surely. \square

Proof This is a direct result of Lemma 4 and Lemma 8. \square

We are now ready to prove Theorem 5 as follows.

Proof (Theorem 5) Given $\mathbf{W}(t_k)$, for any $\tau \geq T_s$, define $\Delta \widetilde{\mathbf{W}}(t_k + \tau) := \mathbf{W}(t_k + \tau) - (\mathbf{W}(t_k) + \tau \mathbf{1} - (\tau - T_s)(\boldsymbol{\rho}^{-1} \circ \mathbf{I}(t_k)))$, where $\boldsymbol{\rho}^{-1} := (\rho_1^{-1}, \dots, \rho_N^{-1})$ and $\boldsymbol{\rho}^{-1} \circ \mathbf{I}(t_k)$ denotes the element-wise product (also called Hadamard product) of $\boldsymbol{\rho}^{-1}$ and $\mathbf{I}(t_k)$. Therefore, at time t_{k+1} , under the W-BMW policy, we have

$$\left(1 + \frac{T_s}{G(\mathbf{W}(t_k))}\right) \left(\mathbf{I}(t_k)^T (\mathbf{W}(t_k) + T_k \mathbf{1} - (T_k - T_s)(\boldsymbol{\rho}^{-1} \circ \mathbf{I}(t_k)) + \Delta \widetilde{\mathbf{W}}(t_{k+1}))\right) \quad (64)$$

$$\leq \left(\mathbf{I}(t_{k+1})^T (\mathbf{W}(t_k) + T_k \mathbf{1} + \Delta \widetilde{\mathbf{W}}(t_{k+1}))\right) \quad (65)$$

Since $\mathbf{I}(t_k)^T \mathbf{W}(t_k) \geq \mathbf{I}(t_{k+1})^T \mathbf{W}(t_k)$, we can rearrange (64) and (65) as

$$\frac{T_s \mathbf{I}(t_k)^T \mathbf{W}(t_k)}{G(\mathbf{W}(t_k))} \leq \left(1 + \frac{T_s}{G(\mathbf{W}(t_k))}\right) \left(\left(T_k \mathbf{I}(t_k)^T (\boldsymbol{\rho}^{-1} - \mathbf{1})\right) - \mathbf{I}(t_k)^T \Delta \widetilde{\mathbf{W}}(t_{k+1})\right) \quad (66)$$

$$+ T_k \mathbf{I}(t_{k+1})^T \mathbf{1} + \mathbf{I}(t_{k+1})^T \Delta \widetilde{\mathbf{W}}(t_{k+1}) \quad (67)$$

$$\leq T_k \left(\frac{\mu_{\max}(T_s + 1)}{\lambda_{\min}} + 1\right) + (T_s + 2) \sum_{i=1}^N |\Delta \widetilde{\mathbf{W}}_i(t_{k+1})|. \quad (68)$$

By the Functional Law of Iterated Logarithm [7], with probability one we have

$$\Delta \widetilde{W}_i(t_{k+1}) = O(\sqrt{T_k \log \log T_k}), \quad \forall i = 1, \dots, N. \quad (69)$$

Therefore, we have

$$\left(\sum_{k=1}^{K_{\mathbb{T}}} \left(\mathbf{1}^T \mathbf{W}(t_k) \right)^{1-\alpha} \right) \leq \frac{N}{T_s} \sum_{k=1}^{K_{\mathbb{T}}} \left(\left(\frac{\mu_{\max}(T_s + 1)}{\lambda_{\min}} + 1 \right) T_k + O(\sqrt{T_k \log \log T_k}) \right). \quad (70)$$

By dividing both sides of (70) by $K_{\mathbb{T}}$ and using Theorem 6, there must exist some constant $B_0 < \infty$ such that

$$\lim_{\mathbb{T} \rightarrow \infty} \frac{\sum_{k=1}^{K_{\mathbb{T}}} \left(\mathbf{1}^T \mathbf{W}(t_k) \right)^{1-\alpha}}{K_{\mathbb{T}}} \leq \frac{B_0}{\epsilon^*}. \quad (71)$$

By the Functional Law of Iterated Logarithm, as ϵ^* approaches 0, with probability one we further have

$$Q_i(t_k) = \lambda_i W_i(t_k) + O(\sqrt{W_i(t_k) \log \log W_i(t_k)}), \quad \forall i \in \mathcal{N}. \quad (72)$$

Therefore, from (71) and (72), there exists another constant $B_1 < \infty$ such that

$$\lim_{\mathbb{T} \rightarrow \infty} \frac{\sum_{k=1}^{K_{\mathbb{T}}} \left(\mathbf{1}^T \mathbf{Q}(t_k) \right)^{1-\alpha}}{K_{\mathbb{T}}} \leq \frac{B_1}{\epsilon^*}. \quad (73)$$

For any $\alpha \in (0, 1)$, by using Theorem 4, we know that the Markov chain induced by $\{\mathbf{Q}(t)\}_{t \geq 0}$ is positive recurrent and hence

$$\mathbb{E} \left[\left(\mathbf{1}^T \mathbf{Q}(t_k) \right)^{1-\alpha} \right] = \lim_{\mathbb{T} \rightarrow \infty} \frac{\sum_{k=1}^{K_{\mathbb{T}}} \left(\mathbf{1}^T \mathbf{Q}(t_k) \right)^{1-\alpha}}{K_{\mathbb{T}}} \quad (74)$$

By Lemma 6, we obtain the queue length bound as

$$\lim_{\epsilon^* \downarrow 0} \epsilon^* \mathbb{E} \left[\left(\mathbf{1}^T \mathbf{Q}(t) \right)^{1-\alpha} \right] \leq B, \quad (75)$$

for some finite constant $B > 0$. By choosing the parameter α to be arbitrarily close to 0, the W-BMW scheduling policy can achieve asymptotically tight queue length upper bound and hence is delay-optimal. \square

7 Simulation

In this section, we explore the delay performance of the two types of BMW policies and the state-of-the-art VFMW policy through extensive simulation of the following three applications: polling systems, directional-antenna systems, and traffic control for signalized intersections. Throughout this section, the arrival and service process of each queue i is Bernoulli with mean λ_i and μ_i , respectively.

7.1 Polling Systems With Arbitrary Service Order

We consider a polling system of 4 parallel queues where the service order can be determined dynamically. We first check the delay performance of the BMW policies with different α . Figure 2(a) and 2(b) show the total average queue length under Q-BMW policy in the two scenarios described in Section 3.2. We state the scenarios here again for easy reference.

- Scenario I: $\lambda = (0.119, 0.119, 0.119, 0.119)$, $\mu = (0.5, 0.5, 0.5, 0.5)$
- Scenario II: $\lambda = (0.08, 0.25, 0.09, 0.01)$, $\mu = (0.8, 0.5, 0.3, 0.2)$

As stated in Theorem 2, the Q-BMW policy achieves the smallest average delay when α is arbitrarily close to 0. Similarly, Figure 3(a) and 3(b) show that under the W-BMW policy the average delay is the smallest when α is arbitrarily close to 0. For consistency of simulation results of different scenarios, for the rest of the simulation we choose $\alpha = 0.001$ for both the Q-BMW and W-BMW policies.

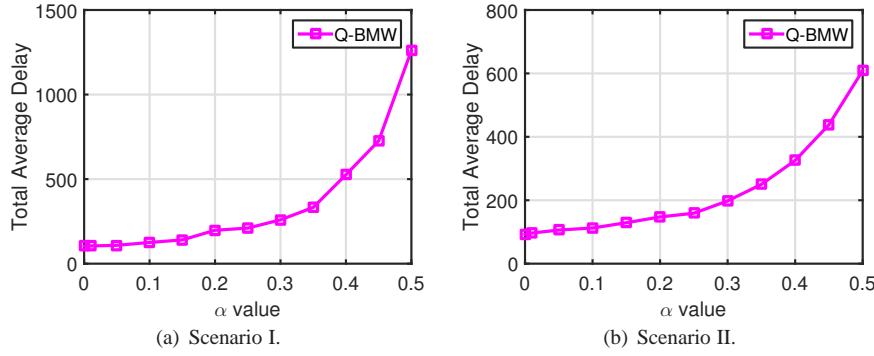


Fig. 2 Average delay versus different α value under Q-BMW policy in Scenario I and II.

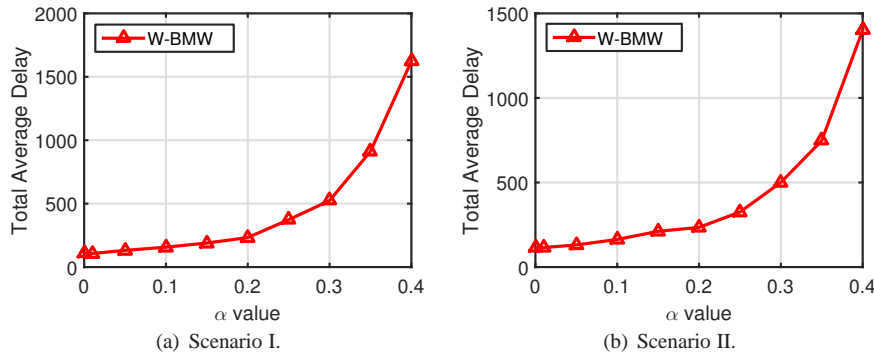


Fig. 3 Average delay versus different α value under W-BMW policy in Scenario I and II.

Next, we simulate the average delay with different utilization factor β^* under the three scheduling policies, as shown in Figure 4(a) and 4(b). We consider both the symmetric and asymmetric cases:

- Scenario III: $\lambda = \beta^* \cdot (0.125, 0.125, 0.125, 0.125)$, $\mu = (0.5, 0.5, 0.5, 0.5)$
- Scenario IV: $\lambda = \beta^* \cdot (0.25, 0.15, 0.075, 0.025)$, $\mu = (0.5, 0.5, 0.5, 0.5)$

In Figure 4(a) and 4(b), we observe that both Q-BMW and W-BMW achieve a much lower average delay than that of the VFMW policy with either frame size function $(\sum_i Q_i(t))^{0.5}$ or $(\sum_i Q_i(t))^{0.99}$. Moreover, we are also interested in the delay performance with different amount of switching overhead. Figure 5(a) and 5(b) show the total average delay with T_s ranging from 1 to 7 under Scenario III and IV with utilization factor equal to 0.95. In these two figures we do not show the simulation results for VFMW with $\alpha = 0.5$ simply because its delay is much larger than its counterparts. We observe that the total average delay grows roughly linearly with the switching overhead and the two BMW policies still have much smaller delay than the VFMW policy, regardless of the amount of switching overhead. Therefore, for the rest of the simulation, we simply choose $T_s = 1$. Figure 6 shows the per-queue average delay of the two BMW policies in Scenario IV. Under the Q-BMW policy, the per-queue delay is inversely proportional to the mean arrival rate. On the other hand, the delay of each queue is about the same under the W-BMW policy. Hence, W-BMW indeed achieves better fairness in per-queue delay than the Q-BMW policy.

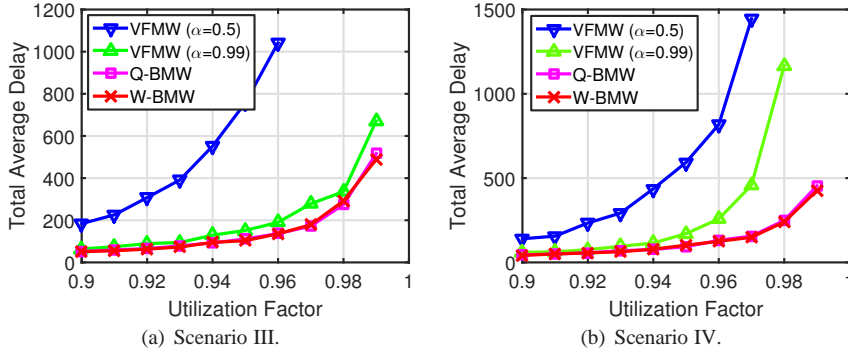


Fig. 4 Total average delay under the Q-BMW, W-BMW, and VFMW policies in Scenario III and IV.

7.2 Multi-Beam Directional-Antenna Systems

We consider a system of 6 queues and $|\mathbf{I}| = 4$ for every feasible schedule \mathbf{I} . In the context of directional-antenna systems, this example represents a 4-beam system. Besides, there are system-wise conflicting constraints which limit the number of maximal feasible schedules. We consider the topology with sets of conflicting queues: queue 1 and queue 2 cannot be served simultaneously; queue 3 and queue 4 cannot

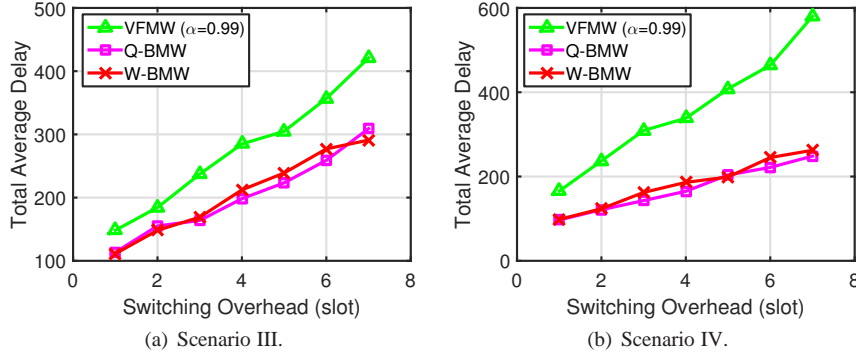


Fig. 5 Total average delay versus different amount of switching overhead in Scenario III and IV.

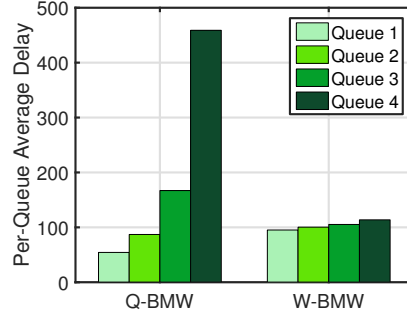


Fig. 6 Per-queue delay under the Q-BMW and W-BMW policies in Scenario IV.

be served simultaneously. Therefore, there are only four maximal feasible schedules: $\{1, 3, 5, 6\}$, $\{1, 4, 5, 6\}$, $\{2, 3, 5, 6\}$, and $\{2, 4, 5, 6\}$. Besides, we choose the traffic pattern to be :

- Scenario V: $\lambda = \beta^* \cdot (0.18, 0.16, 0.25, 0.3, 0.9, 0.8)$ and asymmetric service rates $\mu = (0.3, 0.4, 0.5, 0.6, 0.9, 0.8)$,

with different utilization factor β^* . First, we measure the average delay under the two BMW policies with different α . Figure 7(a) and 7(b) show that the average delay gets lower with smaller α for both Q-BMW and W-BMW policy. This result is consistent with that of the queueing systems where the server can serve at most one queue at a time. Next, with α equal to 0.001, Figure 8(a) shows that the two BMW policies still achieve much smaller delay than that of the VFMW policy. Using the same topology as in Figure 8(a), we change the arrival rates and service rates to be:

- Scenario VI: $\lambda = \beta^* \cdot (0.35, 0.15, 0.3, 0.2, 0.5, 0.5)$ and $\mu_i = 0.5$ for every i .

As shown in Figure 8(b), the result is consistent with that of the other traffic pattern. In summary, with totally different traffic patterns, both Q-BMW and W-BMW always achieve much smaller average delay than that of the VFMW scheduling policy, regardless of the traffic pattern.

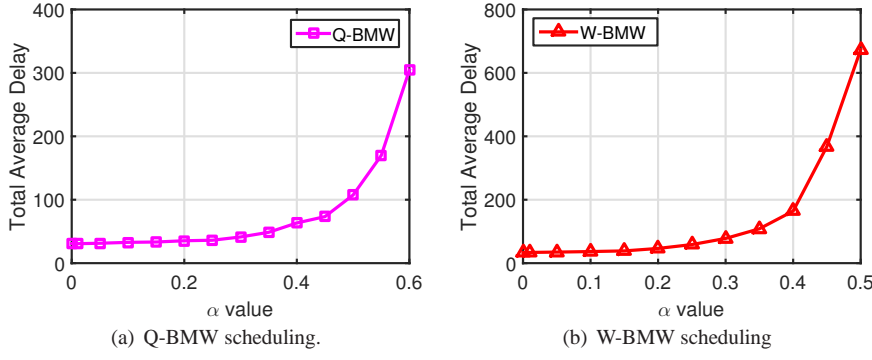


Fig. 7 Average delay under Q-BMW and W-BMW policies with different α in Scenario V.

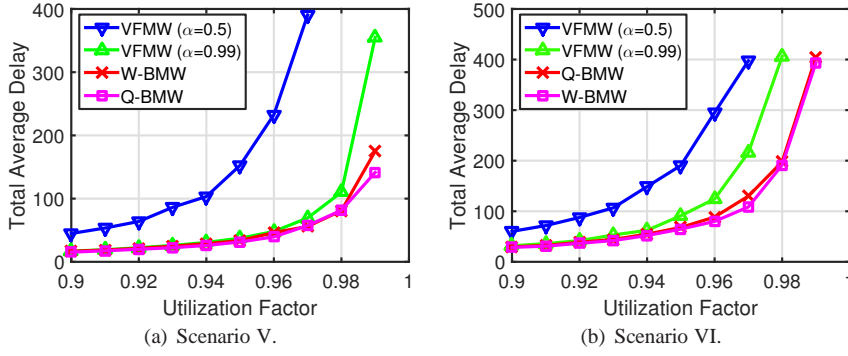


Fig. 8 Delay comparison of the Q-BMW, W-BMW, and VFMW policies in Scenario V and VI.

7.3 Isolated Signalized Intersections

We consider an isolated four-way signalized intersection at which each arriving vehicle either goes straight or makes a left turn. The intersection can be modeled by a queueing system with 8 queues (4 through lanes and 4 left-turn lanes) and $|\mathbf{I}| = 2$ for every feasible schedule \mathbf{I} . Moreover, due to conflicting constraints imposed by the system, there are six maximal feasible schedules. We consider two different traffic patterns as follows:

- Scenario VII: $\lambda = \beta^* \cdot (0.1, 0.5, 0.1, 0.3, 0.1, 0.5, 0.1, 0.3)$ and $\mu_i = 1$ for every queue i .
- Scenario VIII: $\lambda = \beta^* \cdot (0.02, 0.26, 0.24, 0.48, 0.24, 0.48, 0.02, 0.26)$ and $\mu_i = 1$ for every queue i .

Note that the service processes are chosen to be deterministic since the amount of vehicles that are able to pass through the intersection in one time slot should exhibit very little variation. Figure 9(a) and 9(b) show the average delay under the three policies in Scenario VII and VIII. Note that for the VFMW policy we choose $\alpha = 0.8$

instead of $\alpha = 0.99$ simply because the average delay with $\alpha = 0.99$ turns out to be extremely large in these two scenarios. Again, the two BMW policies achieve better system-wise delay performance than the VFMW policy. Besides, note that Figure 9(a) shows that VFMW with $\alpha = 0.5$ performs better than VFMW with $\alpha = 0.8$, while Figure 9(b) shows the opposite. This also highlights the fundamental dilemma of choosing α for the VFMW policy, as discussed in Section 3.2. We further compare the per-queue average delay under Q-BMW and W-BMW. Figure 10(a) and 10(b) show the per-queue delay of queue 1 through queue 4. For simplicity, we do not show the results for queue 5 to queue 8 since they have the same arrival rate pattern and hence have similar per-delay performance as queue 1 to queue 4. These two figures demonstrate that W-BMW still achieves much better fairness than Q-BMW in the sense that the queues with lighter traffic do not suffer from huge queueing delay. Since the per-queue delay is especially crucial in transportation systems, the W-BMW policy is particularly suitable for traffic control at signalized intersections.

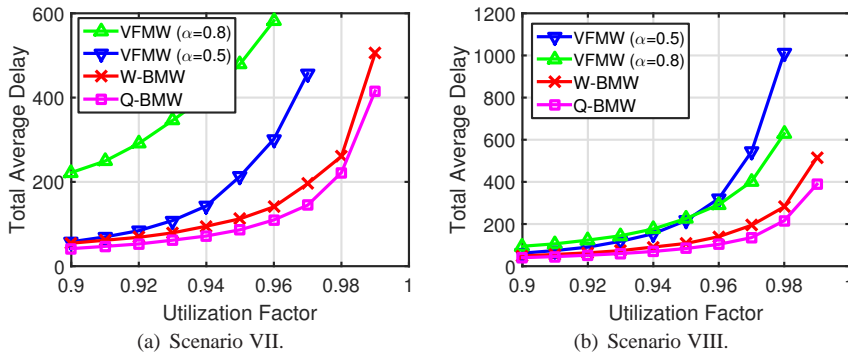


Fig. 9 Total average delay under the Q-BMW, W-BMW, and VFMW policies in Scenario VII and VIII.

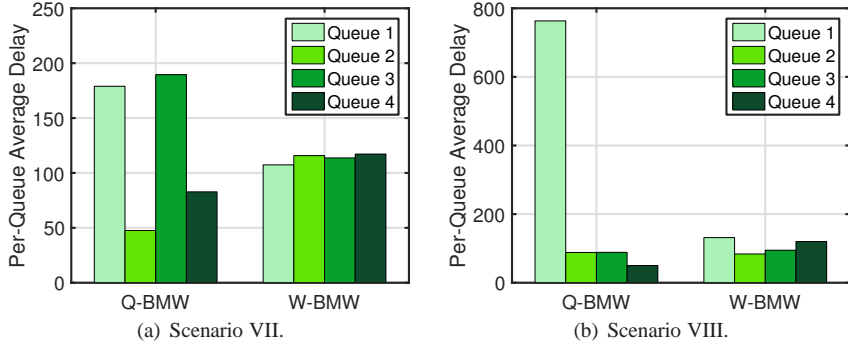


Fig. 10 Per-queue average delay under Q-BMW and W-BMW in Scenario VII and VIII.

8 Conclusion

In this paper, we study the delay performance of queueing systems with switching overhead. We propose two types of BMW scheduling policies that achieves not only throughput-optimality but also delay-optimality. We provide a theoretical queue length upper bound which is asymptotically tight. Through extensive simulation, we demonstrate that the proposed policies achieve much better delay performance than that of the state-of-the-art policy.

Acknowledgements This material is based upon work supported in part by the U. S. Army Research Laboratory and the U. S. Army Research Office under contract/grant number W911NF-15-1-0279 and NPRP Grant 8-1531-2-651 of Qatar National Research Fund (a member of Qatar Foundation).

Appendix 1: Proof of Theorem 1

We prove the theorem by choosing a proper Lyapunov function and showing that the Lyapunov drift is negative. To begin with, consider the multi-step queue length evolution: In the k -th interval, for any τ with $0 < \tau \leq T_k$, for any queue i we have

$$Q_i(t_k + \tau) \leq \max \left\{ 0, Q_i(t_k) - \sum_{s=0}^{\tau-1} M(t_k + s) I_i(t_k + s) S_i(t_k + s) \right\} + \sum_{s=0}^{\tau-1} A_i(t_k + s). \quad (76)$$

Therefore, we also have

$$Q_i(t_k + \tau)^2 \leq \left(Q_i(t_k) - \sum_{s=0}^{\tau-1} M(t_k + s) I_i(t_k + s) S_i(t_k + s) \right)^2 + \left(\sum_{s=0}^{\tau-1} A_i(t_k + s) \right)^2 \quad (77)$$

$$+ 2 \sum_{s=0}^{\tau-1} A_i(t_k + s) \left(Q_i(t_k) - \sum_{s=0}^{\tau-1} M(t_k + s) I_i(t_k + s) S_i(t_k + s) \right) \quad (78)$$

$$\leq Q_i(t_k)^2 + 2Q_i(t_k) \left(\sum_{s=0}^{\tau-1} A_i(t_k + s) - \sum_{s=0}^{\tau-1} M(t_k + s) I_i(t_k + s) S_i(t_k + s) \right) \quad (79)$$

$$+ \left(\sum_{s=0}^{\tau-1} M(t_k + s) I_i(t_k + s) S_i(t_k + s) \right)^2 + \left(\sum_{s=0}^{\tau-1} A_i(t_k + s) \right)^2. \quad (80)$$

Define a Lyapunov function

$$L(t) := \mathbf{Q}(t)^T \mathbf{U} \mathbf{Q}(t), \quad (81)$$

where $\mathbf{U} := \text{diag}(\mu_1^{-1}, \dots, \mu_N^{-1})$. Define $\tilde{F}(\mathbf{Q}(t)) = (\mathbf{1}^T \mathbf{Q}(t))^{\alpha_1}$ with $\alpha_1 \in (0, \alpha)$ and $\alpha + \alpha_1 < 1$. Let $\tilde{T}_k = \min\{T_k, \tilde{F}(\mathbf{Q}(t_k))\}$. For any $\tau_1, \tau_2 \geq 0$, define the

conditional drift between τ_1 and τ_2 as $\Delta(\tau_1, \tau_2) := \mathbb{E}[L(\tau_2) - L(\tau_1) \mid \mathbf{Q}(\tau_1)]$. The conditional drift between t_k and $t_k + \tilde{T}_k$ is

$$\Delta(t_k, t_k + \tilde{T}_k) \quad (82)$$

$$= \mathbb{E} \left[\mathbf{Q}(t_k + \tilde{T}_k)^T \mathbf{U} \mathbf{Q}(t_k + \tilde{T}_k) - \mathbf{Q}(t_k)^T \mathbf{U} \mathbf{Q}(t_k) \mid \mathbf{Q}(t_k) \right] \quad (83)$$

$$\leq 2 \cdot \mathbb{E} \left[\left(\sum_{t=t_k}^{t_k + \tilde{T}_k - 1} (\mathbf{A}(t) - \mathbf{S}^*(t))^T \right) \mathbf{U} \mathbf{Q}(t_k) \mid \mathbf{Q}(t_k) \right] \quad (84)$$

$$+ \mathbb{E} \left[\left(\sum_{t=t_k}^{t_k + \tilde{T}_k - 1} \mathbf{A}(t) \right)^T \mathbf{U} \left(\sum_{t=t_k}^{t_k + \tilde{T}_k - 1} \mathbf{A}(t) \right) \mid \mathbf{Q}(t_k) \right] \quad (85)$$

$$+ \mathbb{E} \left[\left(\sum_{t=t_k}^{t_k + \tilde{T}_k - 1} \mathbf{S}^*(t) \right)^T \mathbf{U} \left(\sum_{t=t_k}^{t_k + \tilde{T}_k - 1} \mathbf{S}^*(t) \right) \mid \mathbf{Q}(t_k) \right], \quad (86)$$

where $\mathbf{S}^*(t) := M(t)(\mathbf{S}(t) \circ \mathbf{I}(t))$ and $\mathbf{S}(t) \circ \mathbf{I}(t)$ denotes the element-wise product (also called Hadamard product) of the two vectors $\mathbf{S}(t)$ and $\mathbf{I}(t)$. For any $t \geq 0$, we have $\mathbf{A}(t) \leq A_{\max} \cdot \mathbf{1}$ and $\mathbf{S}(t) \leq S_{\max} \cdot \mathbf{1}$, regardless of the queue length at time t . Therefore, (85) and (86) are bounded as

$$\mathbb{E} \left[\left(\sum_{t=t_k}^{t_k + \tilde{T}_k - 1} \mathbf{A}(t) \right)^T \mathbf{U} \left(\sum_{t=t_k}^{t_k + \tilde{T}_k - 1} \mathbf{A}(t) \right) \mid \mathbf{Q}(t_k) \right] \leq A_{\max}^2 \text{Tr}(\mathbf{U}) \mathbb{E}[\tilde{T}_k^2 \mid \mathbf{Q}(t_k)] \quad (87)$$

$$\mathbb{E} \left[\left(\sum_{t=t_k}^{t_k + \tilde{T}_k - 1} \mathbf{S}^*(t) \right)^T \mathbf{U} \left(\sum_{t=t_k}^{t_k + \tilde{T}_k - 1} \mathbf{S}^*(t) \right) \mid \mathbf{Q}(t_k) \right] \leq S_{\max}^2 \text{Tr}(\mathbf{U}) \mathbb{E}[\tilde{T}_k^2 \mid \mathbf{Q}(t_k)]. \quad (88)$$

Since both $\mathbf{A}(t)$ and $\mathbf{S}(t)$ are independent of $\mathbf{Q}(t_k)$, we can rewrite (84) as

$$\mathbb{E} \left[\left(\sum_{t=t_k}^{t_k + \tilde{T}_k - 1} (\mathbf{A}(t) - \mathbf{S}^*(t))^T \right) \mathbf{U} \mathbf{Q}(t_k) \mid \mathbf{Q}(t_k) \right] \quad (89)$$

$$= \mathbb{E} \left[(\tilde{T}_k \boldsymbol{\rho}^T - (\tilde{T}_k - T_s) \mathbf{I}(t_k)^T) \mathbf{Q}(t_k) \mid \mathbf{Q}(t_k) \right], \quad (90)$$

where $\boldsymbol{\rho}$ is the vector of normalized traffic load of each queue. Since $\boldsymbol{\lambda}$ is assumed to be in the capacity region, then there exists a J -dimensional non-negative vector $\boldsymbol{\beta}^T = (\beta_1, \dots, \beta_J)$ with $\boldsymbol{\beta}^T \mathbf{1} < 1$ such that $\sum_{j=1}^J \beta_j \mathbf{I}^{(j)} \geq \boldsymbol{\rho}$. Under the Q-BMW

policy, it is guaranteed that $\mathbf{I}(t_k)^T \mathbf{Q}(t_k) = \max_{j: 1 \leq j \leq J} (\mathbf{I}^{(j)})^T \mathbf{Q}(t_k)$. Therefore,

$$\boldsymbol{\rho}^T \mathbf{Q}(t_k) \leq \left(\sum_{j=1}^J \beta_j \mathbf{I}^{(j)} \right)^T \mathbf{Q}(t_k) \quad (91)$$

$$\leq \left(\sum_{j=1}^J \beta_j \right) \mathbf{I}(t_k)^T \mathbf{Q}(t_k) \quad (92)$$

$$= (1 - \epsilon) \mathbf{I}(t_k)^T \mathbf{Q}(t_k) \quad (93)$$

where $\epsilon := 1 - \boldsymbol{\beta}^T \mathbf{1}$ denotes the corresponding "distance" from the boundary of the capacity region. From (85)-(93), the conditional drift can be written as

$$\Delta(t_k, t_k + \tilde{T}_k) \leq \mathbb{E} \left[2 \cdot (-\epsilon \tilde{T}_k + T_s) \mathbf{I}(t_k)^T \mathbf{Q}(t_k) + B_0 \tilde{T}_k^2 \mid \mathbf{Q}(t_k) \right], \quad (94)$$

where $B_0 = (S_{\max}^2 + A_{\max}^2) \text{Tr}(\mathbf{U})$ does not depend on the queue length vector or scheduling decisions. Suppose the server can serve at most K queues at a time. By Lemma 3, we also know that

$$T_k \geq C_0 \left(\mathbf{1}^T \mathbf{Q}(t_k) \right)^{1-\alpha}, \quad (95)$$

where $C_0 = T_s / (NK(A_{\max} + (1 + T_s)S_{\max}))$. Here, we need to discuss two possible cases:

Case 1: $\tilde{F}(\mathbf{Q}(t_k)) \geq C_0 (\mathbf{1}^T \mathbf{Q}(t_k))^{1-\alpha}$

The above condition also implies that

$$\left(\mathbf{1}^T \mathbf{Q}(t_k) \right)^{\alpha_1} \geq C_0 \left(\mathbf{1}^T \mathbf{Q}(t_k) \right)^{1-\alpha}. \quad (96)$$

Therefore, by the assumption that $\alpha + \alpha_1 < 1$, we have

$$\mathbf{1}^T \mathbf{Q}(t_k) \leq C_0^{\frac{-1}{1-\alpha-\alpha_1}}. \quad (97)$$

Hence, from (94) and (97), we know that the conditional drift between t_k and $t_k + \tilde{T}_k$ is bounded, i.e.

$$\Delta(t_k, t_k + \tilde{T}_k) \leq \mathbb{E} \left[2 \cdot \mathbf{I}(t_k)^T \mathbf{Q}(t_k) + B_0 \tilde{T}_k^2 \mid \mathbf{Q}(t_k) \right] \quad (98)$$

$$\leq 2T_s \cdot \mathbf{I}(t_k)^T \mathbf{Q}(t_k) + B_0 \tilde{F}(\mathbf{Q}(t_k))^2 \quad (99)$$

$$\leq 2T_s \cdot \mathbf{1}^T \mathbf{Q}(t_k) + B_0 \left(\mathbf{1}^T \mathbf{Q}(t_k) \right)^{2\alpha_1} \quad (100)$$

$$\leq 2T_s \cdot C_0^{\frac{-1}{1-\alpha-\alpha_1}} + B_0 C_0^{\frac{-2\alpha_1}{1-\alpha-\alpha_1}} < \infty. \quad (101)$$

This also implies that the unconditional drift between t_k and $t_k + \tilde{T}_k$ is bounded, i.e.

$$\mathbb{E} \left[L(t_k + \tilde{T}_k) - L(t_k) \right] \leq 2T_s \cdot C_0^{\frac{-1}{1-\alpha-\alpha_1}} + B_0 C_0^{\frac{-2\alpha_1}{1-\alpha-\alpha_1}} < \infty. \quad (102)$$

Case 2: $\tilde{F}(\mathbf{Q}(t_k)) < C_0(\mathbf{1}^T \mathbf{Q}(t_k))^{1-\alpha}$

The above condition also implies that $\tilde{T}_k = \tilde{F}(\mathbf{Q}(t_k)) < T_k$. Therefore, (94) can then be written as

$$\Delta(t_k, t_k + \tilde{T}_k) \leq 2\left(-\epsilon \tilde{F}(\mathbf{Q}(t_k)) + T_s\right) \mathbf{I}(t_k)^T \mathbf{Q}(t_k) + B_0 \tilde{F}(\mathbf{Q}(t_k))^2 \quad (103)$$

$$\leq -\frac{2\epsilon}{N} \left(\mathbf{1}^T \mathbf{Q}(t_k)\right)^{1+\alpha_1} + 2T_s \left(\mathbf{1}^T \mathbf{Q}(t_k)\right) + B_0 \left(\mathbf{1}^T \mathbf{Q}(t_k)\right)^{2\alpha_1}. \quad (104)$$

Since $-\frac{2\epsilon}{N} \left(\mathbf{1}^T \mathbf{Q}(t_k)\right)^{1+\alpha_1}$ is the dominating term, there must exist some constant $B_2 > 0$ such that

$$\Delta(t_k, t_k + \tilde{T}_k) \leq B_2 - \frac{\epsilon}{N} \left(\mathbf{1}^T \mathbf{Q}(t_k)\right)^{1+\alpha_1}. \quad (105)$$

Moreover, we also know that

$$\sum_{t=t_k}^{t_k+\tilde{T}_k-1} \mathbf{1}^T \mathbf{Q}(t) \leq \sum_{t=t_k}^{t_k+\tilde{T}_k-1} \mathbf{1}^T \left(\mathbf{Q}(t_k) + \sum_{\tau=t_k}^{t_k+\tilde{T}_k-1} \mathbf{A}(\tau) \right) \quad (106)$$

By taking conditional expectation, we have

$$\mathbb{E} \left[\sum_{t=t_k}^{t_k+\tilde{T}_k-1} \mathbf{1}^T \mathbf{Q}(t) \mid \mathbf{Q}(t_k) \right] \leq \mathbb{E} \left[\tilde{T}_k \cdot \mathbf{1}^T \mathbf{Q}(t_k) + \tilde{T}_k^2 \cdot \mathbf{1}^T \boldsymbol{\lambda} \mid \mathbf{Q}(t_k) \right] \quad (107)$$

$$= \left(\mathbf{1}^T \mathbf{Q}(t_k)\right)^{1+\alpha_1} + \mathbf{1}^T \boldsymbol{\lambda} \cdot \left(\mathbf{1}^T \mathbf{Q}(t_k)\right)^{2\alpha_1} \quad (108)$$

$$\leq (1 + \mathbf{1}^T \boldsymbol{\lambda}) \left(\mathbf{1}^T \mathbf{Q}(t_k)\right)^{1+\alpha_1}. \quad (109)$$

The last inequality holds since $\alpha_1 < 1$. Therefore, based on (105) and (109), we obtain

$$\mathbb{E} \left[L(t_k + \tilde{T}_k) - L(t_k) \right] \leq B_2 - \frac{\epsilon}{N_1} \mathbb{E} \left[\sum_{t=t_k}^{t_k+\tilde{T}_k-1} \mathbf{1}^T \mathbf{Q}(t) \right], \quad (110)$$

where $N_1 := N(1 + \mathbf{1}^T \boldsymbol{\lambda})$.

Next, we consider the slot-by-slot conditional drift for any t between $t_k + \tilde{T}_k$ and t_{k+1} . Note that there is no switching between $t_k + \tilde{T}_k$ and t_{k+1} and hence $M(t) = 1$ for all $t \in [t_k + \tilde{T}_k, t_{k+1})$. Therefore,

$$\Delta(t, t+1) = \mathbb{E} \left[\mathbf{Q}(t+1)^T \mathbf{U} \mathbf{Q}(t+1) - \mathbf{Q}(t)^T \mathbf{U} \mathbf{Q}(t) \mid \mathbf{Q}(t) \right] \quad (111)$$

$$\leq 2 \cdot \mathbb{E} \left[(\mathbf{A}(t) - \mathbf{S}^*(t))^T \mathbf{U} \mathbf{Q}(t) \mid \mathbf{Q}(t) \right] \quad (112)$$

$$+ \mathbb{E} \left[\mathbf{A}(t)^T \mathbf{U} \mathbf{A}(t) + \mathbf{S}^*(t)^T \mathbf{U} \mathbf{S}^*(t) \mid \mathbf{Q}(t) \right] \quad (113)$$

Similar to (87) and (88), we know

$$\mathbb{E} \left[\mathbf{A}(t)^T \mathbf{U} \mathbf{A}(t) + \mathbf{S}^*(t)^T \mathbf{U} \mathbf{S}^*(t) \right] \leq (A_{\max}^2 + S_{\max}^2) \text{Tr}(\mathbf{U}) = B_0. \quad (114)$$

Besides, since $\mathbf{A}(t)$ and $\mathbf{S}(t)$ are independent of $\mathbf{Q}(t)$, (112) can be written as

$$\mathbb{E} \left[(\mathbf{A}(t) - \mathbf{S}^*(t))^T \mathbf{U} \mathbf{Q}(t) \mid \mathbf{Q}(t) \right] = (\boldsymbol{\rho}^T - \mathbf{I}(t)^T) \mathbf{Q}(t) \quad (115)$$

Hence, we have

$$\Delta(t, t+1) \leq 2 \cdot (\boldsymbol{\rho}^T - \mathbf{I}(t)^T) \mathbf{Q}(t) + B_0 \quad (116)$$

Under the Q-BMW policy, at time t we must have

$$\mathbf{I}(t)^T \mathbf{Q}(t) \geq (\mathbf{I}^{(j)})^T \mathbf{Q}(t) - \frac{\mathbf{I}(t)^T \mathbf{Q}(t)}{F(\mathbf{Q}(t_k))}, \quad \forall j = 1, \dots, N \quad (117)$$

Along with (91)-(93), we then have

$$(1 - \epsilon) \mathbf{I}(t)^T \mathbf{Q}(t) = \sum_{j=1}^J \beta_j \mathbf{I}(t)^T \mathbf{Q}(t) \quad (118)$$

$$\geq \sum_{j=1}^J \beta_j (\mathbf{I}^{(j)})^T \mathbf{Q}(t) - \sum_{j=1}^J \beta_j \frac{\mathbf{I}^T(t) \mathbf{Q}(t)}{F(\mathbf{Q}(t_k))} \quad (119)$$

$$\geq \boldsymbol{\rho}^T \mathbf{Q}(t) - (1 - \epsilon) \frac{\mathbf{I}^T(t) \mathbf{Q}(t)}{F(\mathbf{Q}(t_k))}. \quad (120)$$

Therefore, (116) can be written as

$$\Delta(t, t+1) \leq 2 \cdot \left(-\epsilon \mathbf{I}(t)^T \mathbf{Q}(t) + (1 - \epsilon) \frac{\mathbf{I}(t)^T \mathbf{Q}(t)}{F(\mathbf{Q}(t_k))} \right) + B_0 \quad (121)$$

$$\leq 2 \cdot \left(-\frac{\epsilon}{N} (\mathbf{1}(t)^T \mathbf{Q}(t)) + (1 - \epsilon) \frac{\mathbf{I}(t)^T \mathbf{Q}(t)}{F(\mathbf{Q}(t_k))} \right) + B_0 \quad (122)$$

$$\leq 2 \cdot \left(-\frac{\epsilon}{N} (\mathbf{1}(t)^T \mathbf{Q}(t)) + (1 - \epsilon) (\mathbf{I}(t)^T \mathbf{Q}(t))^{1-\alpha} \right) + B_0 \quad (123)$$

$$\leq 2 \cdot \left(-\frac{\epsilon}{N} (\mathbf{1}(t)^T \mathbf{Q}(t)) + (1 - \epsilon) (\mathbf{1}(t)^T \mathbf{Q}(t))^{1-\alpha} \right) + B_0 \quad (124)$$

Since $\alpha > 0$, then $-\frac{\epsilon}{N} \mathbf{I}(t)^T \mathbf{Q}(t)$ is the dominating term in (124). In other words, there must exist some constant $B_3 > 0$ such that

$$\Delta(t, t+1) \leq B_3 - \frac{\epsilon}{N} (\mathbf{1}(t)^T \mathbf{Q}(t)). \quad (125)$$

Hence, for any $t \in (t_k + \tilde{T}_k, t_{k+1})$, we know

$$\mathbb{E} [L(t+1) - L(t)] \leq B_3 - \frac{\epsilon}{N} \mathbb{E} [\mathbf{1}(t)^T \mathbf{Q}(t)]. \quad (126)$$

Now, we consider any large \mathbb{T} and let $K_{\mathbb{T}}$ be the number of intervals in $[0, \mathbb{T})$. Since each interval lasts for at least one slot, then $K_{\mathbb{T}} \leq \mathbb{T}$. The unconditional drift in $[0, \mathbb{T})$

$$\mathbb{E}[L(\mathbb{T}) - L(0)] = \sum_{k=0}^{K_{\mathbb{T}}-1} \mathbb{E}[L(t_{k+1}) - L(t_k)] \quad (127)$$

$$= \sum_{k=0}^{K_{\mathbb{T}}-1} \left(\mathbb{E}[L(t_k + \tilde{T}_k) - L(t_k)] + \sum_{\tau=t_k+\tilde{T}_k}^{t_{k+1}-1} \mathbb{E}[L(\tau+1) - L(\tau)] \right) \quad (128)$$

$$\leq K_{\mathbb{T}} B_2 - \frac{\epsilon}{N_1} \sum_{k=1}^{K_{\mathbb{T}}-1} \left(\mathbb{E} \left[\sum_{t=t_k}^{t_k+\tilde{T}_k-1} \mathbf{1}^T \mathbf{Q}(t) \right] \right) \quad (129)$$

$$+ B_3 \mathbb{T} - \frac{\epsilon}{N} \sum_{k=0}^{K_{\mathbb{T}}-1} \left(\mathbb{E} \left[\sum_{\tau=t_k+\tilde{T}_k}^{t_{k+1}-1} \mathbf{1}^T \mathbf{Q}(t) \right] \right) \quad (130)$$

$$\leq K_{\mathbb{T}} B_2 + B_3 \mathbb{T} - \frac{\epsilon}{N_1} \left(\mathbb{E} \left[\sum_{t=0}^{\mathbb{T}-1} \mathbf{1}^T \mathbf{Q}(t) \right] \right). \quad (131)$$

Since $L(0) = 0$ and $L(t)$ is nonnegative regardless of t , by letting $\mathbb{T} \rightarrow \infty$, we have

$$\lim_{\mathbb{T} \rightarrow \infty} \frac{\mathbb{E} \left[\sum_{t=0}^{\mathbb{T}-1} \mathbf{1}^T \mathbf{Q}(t) \right]}{\mathbb{T}} \leq \frac{N_1(B_2 + B_3)}{\epsilon} < \infty. \quad (132)$$

□

Appendix 2: Proof of Theorem 4

To begin with, we describe how the HOL waiting time evolves. For each queue i , define $\delta_i(t) := I_i(t) \sum_{m=1}^{S_i(t)} V_i(\varphi_i(t) + m)$. Note that $\delta_i(t)$ represents the potential decrease in HOL waiting time due to the potential service of queue i at time t . We use the boldface symbol $\boldsymbol{\delta}$ to denote the N -dimensional vector $(\delta_1, \dots, \delta_N) \in \mathbb{N}_0^N$. In the k -th interval, for any τ with $0 < \tau \leq T_k$, for any queue i we have

$$W_i(t_k + \tau) \leq \tau + \max \left\{ 0, W_i(t_k) - \sum_{s=0}^{\tau-1} M(t_k + s) \delta_i(t_k + s) \right\}. \quad (133)$$

Recall that $M(t)$ represents whether the server is in ACTIVE mode at time t . Therefore, we also have

$$W_i(t_k + \tau)^2 \leq W_i(t_k)^2 - 2W_i(t_k) \left(\sum_{s=0}^{\tau-1} M(t_k + s) \delta_i(t_k + s) \right) \quad (134)$$

$$+ \left(\sum_{s=0}^{\tau-1} M(t_k + s) \delta_i(t_k + s) \right)^2 + 2\tau W_i(t_k) + \tau^2. \quad (135)$$

Define a Lyapunov function

$$L_2(t) := \mathbf{W}(t)^T \mathbf{P} \mathbf{W}(t), \quad (136)$$

where $\mathbf{P} := \text{diag}(\rho_1, \dots, \rho_N)$. Define $\hat{G}(\mathbf{W}(t)) = (\mathbf{1}^T \mathbf{W}(t))^{\alpha_1}$ with $\alpha_1 \in (0, \alpha)$ and $\alpha + \alpha_1 < 1$. Let $\hat{T}_k = \min\{T_k, \hat{G}(\mathbf{W}(t_k))\}$. For any $\tau_1, \tau_2 \geq 0$, define the conditional drift between τ_1 and τ_2 as $\Delta L_2(\tau_1, \tau_2) := \mathbb{E}[L_2(\tau_2) - L_2(\tau_1) \mid \mathbf{W}(\tau_1)]$. Based on (134) and (135), the conditional drift between t_k and $t_k + \hat{T}_k$ is

$$\Delta L_2(t_k, t_k + \hat{T}_k) \quad (137)$$

$$= \mathbb{E} \left[\mathbf{W}(t_k + \hat{T}_k)^T \mathbf{P} \mathbf{W}(t_k + \hat{T}_k) - \mathbf{W}(t_k)^T \mathbf{P} \mathbf{W}(t_k) \mid \mathbf{W}(t_k) \right] \quad (138)$$

$$\leq -2 \cdot \mathbb{E} \left[\left(\sum_{t=t_k}^{t_k + \hat{T}_k - 1} M(t) \delta(t) \right)^T \mathbf{P} \mathbf{W}(t_k) \mid \mathbf{W}(t_k) \right] \quad (139)$$

$$+ \mathbb{E} \left[\left(\sum_{t=t_k}^{t_k + \hat{T}_k - 1} M(t) \delta(t) \right)^T \mathbf{P} \left(\sum_{t=t_k}^{t_k + \hat{T}_k - 1} M(t) \delta(t) \right) \mid \mathbf{W}(t_k) \right] \quad (140)$$

$$+ 2 \cdot \mathbb{E} \left[(\hat{T}_k \mathbf{1})^T \mathbf{P} \mathbf{W}(t_k) \mid \mathbf{W}(t_k) \right] + \mathbb{E} \left[\hat{T}_k^2 \boldsymbol{\rho}^T \mathbf{1} \mid \mathbf{W}(t_k) \right]. \quad (141)$$

First, we know

$$\mathbb{E} \left[(\hat{T}_k \mathbf{1})^T \mathbf{P} \mathbf{W}(t_k) \mid \mathbf{W}(t_k) \right] = \mathbb{E} \left[\hat{T}_k \boldsymbol{\rho}^T \mathbf{W}(t_k) \mid \mathbf{W}(t_k) \right]. \quad (142)$$

For any $t \geq 0$, by the assumptions on inter-arrival times and service processes, we have $V_i(m) \leq V_{\max}$ for every queue i and $m \geq 0$, and $\mathbf{S}(t) \leq S_{\max} \cdot \mathbf{1}$, regardless of the HOL waiting time at time t . Therefore, (140) is bounded as

$$\mathbb{E} \left[\left(\sum_{t=t_k}^{t_k + \hat{T}_k - 1} M(t) \delta(t) \right)^T \mathbf{P} \left(\sum_{t=t_k}^{t_k + \hat{T}_k - 1} M(t) \delta(t) \right) \mid \mathbf{W}(t_k) \right] \quad (143)$$

$$\leq V_{\max}^2 S_{\max}^2 \text{Tr}(\mathbf{P}) \mathbb{E} \left[\hat{T}_k^2 \mid \mathbf{W}(t_k) \right]. \quad (144)$$

Note that $\mathbf{W}(t_k)$ only tells us when the HOL job arrived and therefore provides no information about the inter-arrival times. Hence, $V_i(m)$ is independent of $\mathbf{W}(t_k)$, for any queue i and $m \geq 0$. Besides, since $\mathbf{S}(t)$ is also independent of the waiting time $\mathbf{W}(t_k)$, we can rewrite (139) as

$$\mathbb{E} \left[\left(\sum_{t=t_k}^{t_k + \hat{T}_k - 1} M(t) \delta(t) \right)^T \mathbf{P} \mathbf{W}(t_k) \mid \mathbf{W}(t_k) \right] \quad (145)$$

$$= \mathbb{E} \left[\left((\hat{T}_k - T_s) (\mathbf{I}(t_k) \circ \boldsymbol{\rho}^{-1}) \right)^T \mathbf{P} \mathbf{W}(t_k) \mid \mathbf{W}(t_k) \right] \quad (146)$$

$$= \mathbb{E} \left[(\hat{T}_k - T_s) \cdot \mathbf{I}(t_k)^T \mathbf{W}(t_k) \mid \mathbf{W}(t_k) \right], \quad (147)$$

where $\boldsymbol{\rho}^{-1} := (\rho_1^{-1}, \dots, \rho_N^{-1})$ is the vector of the reciprocal of per-queue normalized traffic load. Since the system is assumed to be stabilizable, then there exists

a J -dimensional nonnegative vector $\beta^T = (\beta_1, \dots, \beta_J)$ with $\beta^T \mathbf{1} < 1$ such that $\sum_{j=1}^J \beta_j \mathbf{I}^{(j)} \geq \rho$. Under the W-BMW policy, it is guaranteed that $\mathbf{I}(t_k)^T \mathbf{W}(t_k) = \max_{j: 1 \leq j \leq J} (\mathbf{I}^{(j)})^T \mathbf{W}(t_k)$. Therefore, we know

$$\rho^T \mathbf{W}(t_k) \leq \left(\sum_{j=1}^J \beta_j \mathbf{I}^{(j)} \right)^T \mathbf{W}(t_k) \quad (148)$$

$$\leq \left(\sum_{j=1}^J \beta_j \right) \mathbf{I}(t_k)^T \mathbf{W}(t_k) \quad (149)$$

$$= (1 - \epsilon) \mathbf{I}(t_k)^T \mathbf{W}(t_k) \quad (150)$$

where $\epsilon := 1 - \beta^T \mathbf{1}$ denotes the normalized distance between the arrival rate vector and the boundary of the capacity region. From (139)-(150), the conditional drift can be written as

$$\Delta L_2(t_k, t_k + \hat{T}_k) \leq \mathbb{E} \left[2 \cdot (-\epsilon \hat{T}_k + T_s) \mathbf{I}(t_k)^T \mathbf{W}(t_k) + \Phi_0 \hat{T}_k^2 \mid \mathbf{W}(t_k) \right], \quad (151)$$

where $\Phi_0 := (V_{\max}^2 S_{\max}^2 \text{Tr}(\mathbf{P}) + \rho^T \mathbf{1})$ does not depend on the waiting time vector or scheduling decisions. By Lemma 7, we also know that

$$T_k \geq C_1 \left(\mathbf{1}^T \mathbf{W}(t_k) \right)^{1-\alpha}, \quad (152)$$

where $C_1 = T_s / (NK(1 + (1 + T_s)S_{\max}V_{\max}))$. Here, we need to discuss two possible cases:

Case 1: $\hat{G}(\mathbf{W}(t_k)) \geq C_1 \left(\mathbf{1}^T \mathbf{W}(t_k) \right)^{1-\alpha}$

We first have

$$\left(\mathbf{1}^T \mathbf{W}(t_k) \right)^{\alpha_1} \geq C_1 \left(\mathbf{1}^T \mathbf{W}(t_k) \right)^{1-\alpha} \quad (153)$$

Therefore, by the assumption that $\alpha + \alpha_1 < 1$, we have

$$\mathbf{1}^T \mathbf{W}(t_k) \leq C_1^{\frac{-1}{1-\alpha-\alpha_1}} \quad (154)$$

Hence, from (151) and (154), we know that the conditional drift between t_k and $t_k + \hat{T}_k$ is bounded, i.e.

$$\Delta L_2(t_k, t_k + \hat{T}_k) \leq \mathbb{E} \left[2T_s \cdot \mathbf{I}(t_k)^T \mathbf{W}(t_k) + \Phi_0 \hat{T}_k^2 \mid \mathbf{W}(t_k) \right] \quad (155)$$

$$\leq 2T_s \cdot \mathbf{I}(t_k)^T \mathbf{W}(t_k) + \Phi_0 \hat{G}(\mathbf{W}(t_k))^2 \quad (156)$$

$$\leq 2T_s \cdot \mathbf{1}^T \mathbf{W}(t_k) + \Phi_0 \left(\mathbf{1}^T \mathbf{W}(t_k) \right)^{2\alpha_1} \quad (157)$$

$$\leq 2T_s \cdot C_1^{\frac{-1}{1-\alpha-\alpha_1}} + \Phi_0 C_1^{\frac{-2\alpha_1}{1-\alpha-\alpha_1}} < \infty. \quad (158)$$

This also implies that the unconditional drift between t_k and $t_k + \hat{T}_k$ is bounded, i.e.

$$\mathbb{E} \left[L_2(t_k + \hat{T}_k) - L_2(t_k) \right] \leq 2T_s \cdot C_1^{\frac{-1}{1-\alpha-\alpha_1}} + \Phi_0 C_1^{\frac{-2\alpha_1}{1-\alpha-\alpha_1}} < \infty. \quad (159)$$

Case 2: $\hat{G}(\mathbf{W}(t_k)) < C_1 \left(\mathbf{1}^T \mathbf{W}(t_k) \right)^{1-\alpha}$

The above condition implies that $\hat{T}_k = \hat{G}(\mathbf{W}(t_k)) < T_k$. Therefore, (151) can then be written as

$$\Delta L_2(t_k, t_k + \hat{T}_k) \leq 2 \left(-\epsilon \hat{G}(\mathbf{W}(t_k)) + T_s \right) \mathbf{I}(t_k)^T \mathbf{W}(t_k) + \Phi_0 \hat{G}(\mathbf{W}(t_k))^2 \quad (160)$$

$$\leq -\frac{2\epsilon}{N} \left(\mathbf{1}^T \mathbf{W}(t_k) \right)^{1+\alpha_1} + 2T_s \left(\mathbf{1}^T \mathbf{W}(t_k) \right) + \Phi_0 \left(\mathbf{1}^T \mathbf{W}(t_k) \right)^{2\alpha_1}. \quad (161)$$

Since $-\frac{2\epsilon}{N} \left(\mathbf{1}^T \mathbf{W}(t_k) \right)^{1+\alpha_1}$ is the dominating term in (161), there must exist some constant $\Phi_1 > 0$ such that

$$\Delta L_2(t_k, t_k + \hat{T}_k) \leq \Phi_1 - \frac{\epsilon}{N} \left(\mathbf{1}^T \mathbf{W}(t_k) \right)^{1+\alpha_1}. \quad (162)$$

Moreover, we also know that

$$\sum_{t=t_k}^{t_k + \hat{T}_k - 1} \mathbf{1}^T \mathbf{W}(t) \leq \sum_{\tau=0}^{\hat{T}_k - 1} \mathbf{1}^T \left(\mathbf{W}(t_k) + \tau \right) \leq \hat{T}_k \cdot \mathbf{1}^T \mathbf{W}(t_k) + N \hat{T}_k^2. \quad (163)$$

By taking conditional expectation, we have

$$\mathbb{E} \left[\sum_{t=t_k}^{t_k + \hat{T}_k - 1} \mathbf{1}^T \mathbf{W}(t) \mid \mathbf{W}(t_k) \right] \leq \mathbb{E} \left[\hat{T}_k \cdot \mathbf{1}^T \mathbf{W}(t_k) + N \hat{T}_k^2 \mid \mathbf{W}(t_k) \right] \quad (164)$$

$$= \left(\mathbf{1}^T \mathbf{W}(t_k) \right)^{1+\alpha_1} + N \left(\mathbf{1}^T \mathbf{W}(t_k) \right)^{2\alpha_1} \quad (165)$$

$$\leq (1 + N) \left(\mathbf{1}^T \mathbf{W}(t_k) \right)^{1+\alpha_1}. \quad (166)$$

The last inequality holds since $\alpha_1 < 1$. Therefore, based on (162) and (166), we obtain that

$$\mathbb{E} \left[L_2(t_k + \hat{T}_k) - L_2(t_k) \right] \leq \Phi_1 - \frac{\epsilon}{N_2} \mathbb{E} \left[\sum_{t=t_k}^{t_k + \hat{T}_k - 1} \mathbf{1}^T \mathbf{W}(t) \right], \quad (167)$$

where $N_2 := N(N + 1)$.

Next, we consider the slot-by-slot conditional drift for any t between $t_k + \hat{T}_k$ and t_{k+1} . Note that there is no switching between $t_k + \hat{T}_k$ and t_{k+1} . Therefore,

$$\Delta L_2(t, t+1) = \mathbb{E} \left[\mathbf{W}(t+1)^T \mathbf{P} \mathbf{W}(t+1) - \mathbf{W}(t)^T \mathbf{P} \mathbf{W}(t) \mid \mathbf{W}(t) \right] \quad (168)$$

$$\leq 2 \cdot \mathbb{E} \left[\left(\mathbf{1} - \delta(t) \right)^T \mathbf{P} \mathbf{W}(t) \mid \mathbf{W}(t) \right] \quad (169)$$

$$+ \mathbb{E} \left[\left(\mathbf{1} - \delta(t) \right)^T \mathbf{P} \left(\mathbf{1} - \delta(t) \right) \mid \mathbf{W}(t) \right], \quad (170)$$

Similar to (143) and (144), we know

$$\mathbb{E} \left[\left(\mathbf{1} - \delta(t) \right)^T \mathbf{P} \left(\mathbf{1} - \delta(t) \right) \mid \mathbf{W}(t) \right] \leq (S_{\max}^2 V_{\max}^2 + 1) \text{Tr}(\mathbf{P}) = \Phi_0. \quad (171)$$

Besides, since $\mathbf{V}(t)$ and $\mathbf{S}(t)$ are independent of $\mathbf{W}(t)$, (169) can be written as

$$\mathbb{E} \left[\left(\mathbf{1} - \delta(t) \right)^T \mathbf{P} \mathbf{W}(t) \mid \mathbf{W}(t) \right] = (\boldsymbol{\rho}^T - \mathbf{I}(t)^T) \mathbf{W}(t). \quad (172)$$

Hence, we have

$$\Delta L_2(t, t+1) \leq 2 \cdot (\boldsymbol{\rho}^T - \mathbf{I}(t)^T) \mathbf{W}(t) + \Phi_0. \quad (173)$$

Under the W-BMW policy, at time t we must have

$$\mathbf{I}(t)^T \mathbf{W}(t) \geq (\mathbf{I}^{(j)})^T \mathbf{W}(t) - \frac{\mathbf{I}(t)^T \mathbf{W}(t)}{G(\mathbf{W}(t_k))}, \quad \forall j = 1, \dots, N \quad (174)$$

Along with (148)-(150), we then have

$$(1 - \epsilon) \mathbf{I}(t)^T \mathbf{W}(t) = \sum_{j=1}^J \beta_j \mathbf{I}^{(j)}{}^T \mathbf{W}(t) \quad (175)$$

$$\geq \sum_{j=1}^J \beta_j (\mathbf{I}^{(j)})^T \mathbf{W}(t) - \sum_{j=1}^J \beta_j \frac{\mathbf{I}^{(j)}{}^T \mathbf{W}(t)}{G(\mathbf{W}(t_k))} \quad (176)$$

$$\geq \boldsymbol{\rho}^T \mathbf{W}(t) - (1 - \epsilon) \frac{\mathbf{I}(t)^T \mathbf{W}(t)}{G(\mathbf{W}(t_k))}. \quad (177)$$

Therefore, (173) can be written as

$$\Delta L_2(t, t+1) \leq 2 \cdot \left(-\epsilon \mathbf{I}(t)^T \mathbf{W}(t) + (1 - \epsilon) \frac{\mathbf{I}(t)^T \mathbf{W}(t)}{G(\mathbf{W}(t_k))} \right) + \Phi_0 \quad (178)$$

$$\leq 2 \cdot \left(-\frac{\epsilon}{N} (\mathbf{1}(t)^T \mathbf{W}(t)) + (1 - \epsilon) \frac{\mathbf{I}(t)^T \mathbf{W}(t)}{G(\mathbf{W}(t_k))} \right) + \Phi_0 \quad (179)$$

$$\leq 2 \cdot \left(-\frac{\epsilon}{N} (\mathbf{1}(t)^T \mathbf{W}(t)) + (1 - \epsilon) \left(\mathbf{I}(t)^T \mathbf{W}(t) \right)^{1-\alpha} \right) + \Phi_0 \quad (180)$$

$$\leq 2 \cdot \left(-\frac{\epsilon}{N} (\mathbf{1}(t)^T \mathbf{W}(t)) + (1 - \epsilon) \left(\mathbf{1}(t)^T \mathbf{W}(t) \right)^{1-\alpha} \right) + \Phi_0 \quad (181)$$

Since $\alpha > 0$, then $-\frac{\epsilon}{N} \mathbf{1}(t)^T \mathbf{W}(t)$ is the dominating term in (181). In other words, there must exist some constant $\Phi_2 > 0$ such that

$$\Delta L_2(t, t+1) \leq \Phi_2 - \frac{\epsilon}{N} \left(\mathbf{1}(t)^T \mathbf{W}(t) \right). \quad (182)$$

Hence, for any $t \in (t_k + \hat{T}_k, t_{k+1})$, we know

$$\mathbb{E} [L(t+1) - L(t)] \leq \Phi_2 - \frac{\epsilon}{N} \mathbb{E} [\mathbf{1}(t)^T \mathbf{W}(t)]. \quad (183)$$

Now, we consider any large \mathbb{T} and let $K_{\mathbb{T}}$ be the number of intervals in $[0, \mathbb{T})$. Since each interval lasts for at least one slot, then $K_{\mathbb{T}} \leq \mathbb{T}$. The unconditional drift in $[0, \mathbb{T})$

$$\mathbb{E} [L_2(\mathbb{T}) - L_2(0)] = \sum_{k=0}^{K_{\mathbb{T}}-1} \mathbb{E} [L_2(t_{k+1}) - L_2(t_k)] \quad (184)$$

$$= \sum_{k=0}^{K_{\mathbb{T}}-1} \left(\mathbb{E} [L_2(t_k + \hat{T}_k) - L_2(t_k)] + \sum_{\tau=t_k+\hat{T}_k}^{t_{k+1}-1} \mathbb{E} [L_2(\tau+1) - L_2(\tau)] \right) \quad (185)$$

$$\leq K_{\mathbb{T}} \Phi_1 - \frac{\epsilon}{N_2} \sum_{k=1}^{K_{\mathbb{T}}-1} \left(\mathbb{E} \left[\sum_{t=t_k}^{t_k+\hat{T}_k-1} \mathbf{1}^T \mathbf{W}(t) \right] \right) \quad (186)$$

$$+ \Phi_2 \mathbb{T} - \frac{\epsilon}{N} \sum_{k=0}^{K_{\mathbb{T}}-1} \left(\mathbb{E} \left[\sum_{\tau=t_k+\hat{T}_k}^{t_{k+1}-1} \mathbf{1}^T \mathbf{W}(t) \right] \right) \quad (187)$$

$$\leq K_{\mathbb{T}} \Phi_1 + \Phi_2 \mathbb{T} - \frac{\epsilon}{N_2} \left(\mathbb{E} \left[\sum_{t=0}^{\mathbb{T}-1} \mathbf{1}^T \mathbf{W}(t) \right] \right). \quad (188)$$

Since $L_2(0) = 0$ and $L_2(t)$ is nonnegative regardless of t , by letting $\mathbb{T} \rightarrow \infty$, we have

$$\lim_{\mathbb{T} \rightarrow \infty} \frac{\mathbb{E} \left[\sum_{t=0}^{\mathbb{T}-1} \mathbf{1}^T \mathbf{W}(t) \right]}{\mathbb{T}} \leq \frac{N_2(\Phi_1 + \Phi_2)}{\epsilon} \quad (189)$$

Moreover, for any queue i at time t , given the information of $W_i(t)$, we also know

$$\mathbb{E} [Q_i(t) \mid W_i(t)] = \begin{cases} \lambda_i W_i(t) + 1 & , \quad W_i(t) > 0 \\ 0 & , \quad W_i(t) = 0 \end{cases} \quad (190)$$

By taking the unconditional expectation of (190), for any t we have

$$\mathbb{E} [Q_i(t)] \leq \lambda_i \mathbb{E} [W_i(t)] + 1, \quad \forall i \in \mathcal{N}. \quad (191)$$

Hence, we can conclude that

$$\lim_{\mathbb{T} \rightarrow \infty} \frac{\mathbb{E} \left[\sum_{t=0}^{\mathbb{T}-1} \mathbf{1}^T \mathbf{Q}(t) \right]}{\mathbb{T}} \leq \lambda_{\max} \left(\frac{N_2(\Phi_1 + \Phi_2)}{\epsilon} \right) + N < \infty. \quad (192)$$

Hence, the system is strongly stable under the W-BMW policy. \square

References

1. A. Al Hanbali, R. de Haan, R. J. Boucherie, and J.-K. van Ommeren. A tandem queueing model for delay analysis in disconnected ad hoc networks. In *International Conference on Analytical and Stochastic Modeling Techniques and Applications*, pages 189–205, 2008.
2. R. E. Allsop. Estimating the traffic capacity of a signalized road junction. *Transportation Research*, 6(3):245–255, 1972.
3. M. Armony and N. Bambos. Queueing dynamics and maximal throughput scheduling in switched processing systems. *Queueing systems*, 44(3):209–252, 2003.
4. G. Celik, S. C. Borst, P. A. Whiting, and E. Modiano. Dynamic Scheduling with Reconfiguration Delays. *Queueing Syst. Theory Appl.*, 83(1-2):87–129, Jun 2016.
5. G. D. Celik and E. Modiano. Scheduling in networks with time-varying channels and reconfiguration delay. In *Proc. IEEE INFOCOM*, pages 990–998, March 2012.
6. C. W. Chan, M. Armony, and N. Bambos. Maximum Weight Matching with Hysteresis in Overloaded Queues with Setups. *Queueing Syst. Theory Appl.*, 82(3-4):315–351, Apr. 2016.
7. H. Chen and D. D. Yao. *Fundamentals of queueing networks: Performance, asymptotics, and optimization*, volume 46. Springer, 2001.
8. F. M. David, J. C. Carlyle, and R. H. Campbell. Context Switch Overheads for Linux on ARM Platforms. In *Proceedings of the 2007 Workshop on Experimental Computer Science, ExpCS '07*, 2007.
9. A. Eryilmaz and R. Srikant. Asymptotically tight steady-state queue length bounds implied by drift conditions. *Queueing Systems*, 72(3-4):311–359, 2012.
10. A. Eryilmaz, R. Srikant, and J. R. Perkins. Stable scheduling policies for fading wireless channels. *IEEE/ACM Transactions on Networking*, 13(2):411–424, April 2005.
11. Z. Fan. Wireless networking with directional antennas for 60 GHz systems. In *14th European Wireless Conference*, pages 1–7, June 2008.
12. A. Ghavami, K. Kar, and S. Ukkusuri. Delay analysis of signal control policies for an isolated intersection. In *15th International IEEE Conference on Intelligent Transportation Systems*, pages 397–402, 2012.
13. G. R. Gupta and N. B. Shroff. Delay analysis for wireless networks with single hop traffic and general interference constraints. *IEEE/ACM Transactions on Networking*, 18(2):393–405, 2010.
14. Y.-C. Hung and C.-C. Chang. Dynamic scheduling for switched processing systems with substantial service-mode switching times. *Queueing systems*, 60(1-2):87–109, 2008.
15. K. Kar, S. Sarkar, A. Ghavami, and X. Luo. Delay guarantees for throughput-optimal wireless link scheduling. *IEEE Transactions on Automatic Control*, 57(11):2906–2911, 2012.
16. L. B. Le, K. Jagannathan, and E. Modiano. Delay analysis of maximum weight scheduling in wireless ad hoc networks. In *43rd Annual Conference on Information Sciences and Systems (CISS)*, pages 389–394, 2009.
17. S. Le Vine, A. Zolfaghari, and J. Polak. Autonomous cars: the tension between occupant experience and intersection capacity. *Transportation Research Part C: Emerging Technologies*, 52:1–14, 2015.
18. H. Levy and M. Sidi. Polling systems: applications, modeling, and optimization. *IEEE Transactions on Communications*, 38(10):1750–1760, Oct 1990.
19. X. Liu, K. Ma, and P. R. Kumar. Towards provably safe mixed transportation systems with human-driven and automated vehicles. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 4688–4694, Dec 2015.
20. M. P. McGarry, M. Reisslein, and M. Maier. Ethernet passive optical network architectures and dynamic bandwidth allocation algorithms. *IEEE Communications Surveys Tutorials*, 10(3):46–60, 2008.
21. V. Navda, A. P. Subramanian, K. Dhanasekaran, A. Timm-Giel, and S. Das. Mobisteer: Using steerable beam directional antenna for vehicular network access. In *Proceedings of the 5th International Conference on Mobile Systems, Applications and Services, MobiSys '07*, pages 192–205, 2007.
22. M. J. Neely. Delay analysis for maximal scheduling with flow control in wireless networks with bursty traffic. *IEEE/ACM Transactions on Networking*, 17(4):1146–1159, 2009.
23. M. J. Neely. Stability and capacity regions or discrete time queueing networks. *arXiv preprint arXiv:1003.3396*, 2010.
24. T. Nitsche, C. Cordeiro, A. B. Flores, E. W. Knightly, E. Perahia, and J. C. Widmer. IEEE 802.11ad: directional 60 GHz communication for multi-Gigabit-per-second Wi-Fi [Invited Paper]. *IEEE Communications Magazine*, 52(12):132–141, December 2014.

25. Y. Niu, Y. Li, D. Jin, L. Su, and A. V. Vasilakos. A survey of millimeter wave communications (mmWave) for 5G: opportunities and challenges. *Wireless Networks*, 21(8):2657–2676, 2015.
26. J. R. Perkins and P. R. Kumar. Stable, distributed, real-time scheduling of flexible manufacturing/assembly/diassembly systems. *IEEE Transactions on Automatic Control*, 34(2):139–148, Feb 1989.
27. A. Sharifnia, M. Caramanis, and S. B. Gershwin. Dynamic setup scheduling and flow control in manufacturing systems. *Discrete Event Dynamic Systems*, 1(2):149–175, 1991.
28. S. Singh, R. Mudumbai, and U. Madhow. Interference Analysis for Highly Directional 60-GHz Mesh Networks: The Case for Rethinking Medium Access Control. *IEEE/ACM Transactions on Networking*, 19(5):1513–1527, Oct 2011.
29. H. Takagi. Queueing Analysis of Polling Models. *ACM Comput. Surv.*, 20(1):5–28, Mar 1988.
30. H. Takagi. Queueing analysis of polling models: progress in 1990-1994. *Frontiers In Queueing: Models and applications in science and engineering*, 7:119, 1997.
31. L. Tassiulas and A. Ephremides. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Transactions on Automatic Control*, 37(12):1936–1948, Dec 1992.
32. L. Tassiulas and A. Ephremides. Dynamic server allocation to parallel queues with randomly varying connectivity. *IEEE Transactions on Information Theory*, 39(2):466–478, Mar 1993.
33. P. Varaiya. Max pressure control of a network of signalized intersections. *Transportation Research Part C: Emerging Technologies*, 36:177–195, 2013.
34. V. Vishnevskii and O. Semenova. Mathematical methods to study the polling systems. *Automation and Remote Control*, 67(2):173–220, 2006.
35. T. Wongpiromsarn, T. Uthacharoenpong, Y. Wang, E. Frazzoli, and D. Wang. Distributed traffic signal control for maximum network throughput. In *Proc. IEEE Conference on Intelligent Transportation Systems*, pages 588–595, Sept 2012.